DOCUMENT RESUME

ED 268 135 TM 850 736

AUTHOR	Carlson, Sybil B.; And Others
TITLE	Relationship of Admission Test Scores to Writing Performance of Native and Nonnative Speakers of English.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-GRE-83-R2; ETS-RR-85-21; TOEFL-RR-19
PUB DATE	Aug 85
NOTE	152p.
PUB TYPE	Reports - Research/Technical (143) -- Tests/Evaluation Instruments (160)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Arabic; Chinese; *College Entrance Examinations; Computer Software; Correlation; *English (Second Language); Evaluation Criteria; Evaluation Methods; Factor Structure; Higher Education; Holistic Evaluation; *Interrater Reliability; *Language Tests; Native Speakers; Scoring; Second Language Learning; Spanish; Test Reliability; *Writing Evaluation
IDENTIFIERS	Graduate Record Examinations; *Test of English as a Foreign Language

ABSTRACT

Four writing samples were obtained from 638 foreign college applicants who represented three major foreign language groups (Arabic, Chinese, and Spanish), and from 60 native English speakers. All four were scored holistically, two were also scored for sentence-level and discourse-level skills, and some were scored by the Writer's Workbench computer software and by professors in the fields of engineering and the social sciences. Test of English as a Foreign Language (TOEFL) scores were obtained for all foreign students, and Graduate Record Examinations (GRE) General Test scores for some. A GRE score and a multiple-choice writing test score were obtained for the Americans. Findings included: (1) holistic, discourse-level, and sentence-level scores were so closely related that the holistic score alone should suffice; (2) correlations among writing sample topics were as high across as within topic types; (3) scores of English as a second language raters, English raters, and subject matter raters were all highly correlated, suggesting substantial agreement in their standards; (4) correlations and factor analyses indicated that writing samples and TOEFL scores were highly related, but each also reliably measured an independent aspect; and (5) correlations of holistic writing sample scores with GRE item type scores confirmed a previously reported pattern of relationships. (A four-item writing test and numerous tables are appended). (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED268135

TOEFL

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

Relationship of Admission Test Scores to Writing Performance of Native and Nonnative Speakers of English

Sybil B. Carlson
Brent Bridgeman
Roberta Camp
Janet Waanders

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. Weidenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 850 736

2

Relationship of Admission Test Scores to Writing Performance of Native and Nonnative Speakers of English

Sybil B. Carlson
Brent Bridgeman
Roberta Camp
and
Janet Waanders

Educational Testing Service
Princeton, New Jersey

GRE No. 83-R2

RR-85-21

Copyright © 1985 by Educational Testing Service. All rights reserved.

Unauthorized reproduction in whole or in part is prohibited.

TOEFL is a trademark of Educational Testing Service, registered
in the U.S.A. and in many other countries.

Abstract

Four writing samples were obtained from 638 applicants for admission to U.S. institutions as undergraduates or as graduate students in business, engineering, or social science. The applicants represented three major foreign language groups (Arabic, Chinese, and Spanish), plus a small sample of native English speakers. Two of the writing topics were of the compare and contrast type and the other two involved chart and graph interpretation. The writing samples were scored by 23 readers who are English as a second language specialists and 23 readers who are English writing experts. Each of the four writing samples was scored holistically, and during a separate rating session two of the samples from each student were assigned separate scores for sentence-level and discourse-level skills. Representative subsamples of the papers also were scored descriptively with the Writer's Workbench computer program and by graduate-level subject matter professors in engineering and the social sciences.

In addition to the writing sample scores, TOEFL scores were obtained for all students in the foreign sample. GRE General Test scores were obtained for students in the U.S. sample and for a subsample of students in the foreign sample. Students in the U.S. sample also took a multiple-choice measure of writing ability.

Among the key findings were the following: 1) holistic scores, discourse-level scores, and sentence-level scores were so closely related that the holistic score alone should be sufficient; 2) correlations among topics were as high across topic types as within topic types; 3) scores of ESL raters, English raters, and subject matter raters were all highly correlated, suggesting substantial agreement in the standards used; correlations and factor analyses indicated that scores on the writing samples and TOEFL were highly related, but that each also was reliably measuring some aspect of English language proficiency that was not assessed by the other; and (5) correlations of holistic writing sample scores with scores on item types within the sections of the GRE General Test yielded a pattern of relationships that was consistent with the relationships reported in other GRE studies.

Acknowledgments

A considerable number of individuals generously gave their time, energies, and insights during the implementation of this complex project. Many of these people are listed in Appendix B. Faculty members at several institutions throughout the United States collected the pilot test data for us within a tight schedule, work they contributed because they support the kind of writing assessment research we are doing. These colleagues are: Helen Berezovsky, University of Pennsylvania; Kathleen Mellor, Wichita State University; Patricia Dyer, University of Delaware; Douglas Flahive, Colorado State University; Barbara Gray, Polytechnic Institute of New York; George Gadda, UCLA; and Martha McNamara, University of Akron.

Our colleagues at the University of Delaware have given us invaluable advice and assistance, both in our previous survey of academic writing skills and in this study. In particular, Louis A. Arena, the Director of the Writing Center, and Patricia Dyer, Director of the English Language Institute, have been constant supporters and collaborators. Their staff members who worked with us and encouraged our efforts are: Marsha Peoples, Margaret Hassert, and Rajai Khanji, as well as the ESL and Writing Center faculty.

Colleagues outside of ETS and ETS staff members agreed to participate in a one-day reading of the pilot test topics. The ETS staff readers consisted of Charles Stansfield and Barbara Suomi of the TOEFL test development staff. Readers from outside ETS were Patricia Dyer, Director of the English Language Institute program, and Louis A. Arena, Director of the Writing Center, at the University of Delaware; Helen Berezovsky, Assistant Director of the English Program for Foreign Students at the University of Pennsylvania; and George Gadda, instructor at the UCLA Writing Center. Janet Waanders of the College Board test development staff at ETS served as the Chief Reader. The two project directors and Roberta Camp, College Board test development staff, also participated. Lorraine Simon, ETS research staff, helped to organize the logistics of the reading and worked as an aide during the reading.

The international test administrations were very complicated, but Willem Spits and Judith Boyle of TOEFL program administration at ETS helped us through many potential calamities. At AMIDEAST, Washington, D.C., Kay Hogan and Hank Luehmann greatly facilitated our testing procedures at the Arabic centers. We also are grateful to the many test center supervisors who efficiently collected the data at the international and United States centers.

Also at ETS, Gertrude Conlan, of the College Board test development staff, served as a valuable resource to us during the planning of the scoring sessions and throughout the various facets of the writing sample evaluation. The staff of the Essay Reading Office, particularly Kim Kent, Elizabeth Benyon, Deborah O'Neal, and Phyllis Murphy, contributed much of their time, devotion, and expertise to the writing sample reading weekend. Many others of their staff worked most effectively at the reading weekend and in preparing the final writing sample scores.

Sydell Carlton, of ETS College Board test development staff, assisted us in selecting the most appropriate form of an indirect measure of writing ability for the United States students. Roberta Kline, of the ETS Measurement Research and Services staff, painstakingly recorded data on rosters and maintained a high level of quality control. Michaele Mikovsky, of the ETS Measurement Research and Services staff, contributed her graphic arts expertise to the design of the pilot test and pretest wriitng sample stimuli.

At Colorado State University, Joy Reid, of the ESL Department, has been an inspiration to us. She has collaborated with us on a regular basis throughout the project and also assisted in our understanding of the Writer's Workbench system. Joy arranged to have a sample of the writing samples analyzed by the Writer's Workbench. Roberta Scott, a composition instructor as well, keyed in the verbatim writing samples in record time and with the care that is required of research data.

Don Rock, of the Statistical and Psychometric Research and Services division at ETS, skillfully provided advice on the appropriate data analyses and interpretations. Data analysts in the same division conducted the various analyses: Bruce Kaplan, James Rosso, and Richard Harrison.

Thoughtful reviews of the draft report were provided by Hunter Breland, Gordon Hale, Charles Stansfield, Lawrence Stricker, and Protase Woodford.

Finally, we thank the many other individuals who gave their time, efforts, and encouragement.

Table of Contents

Relationship of Admission Test Scores to Writing Performance of Native and Nonnative Speakers of English

Page

I. INTRODUCTION 1

Foundations for the Design and Implementation of the Study--Research, Theory, and Practice 3

A Definition of Writing Competence 4

The New Paradigm for Writing Instruction and Assessment 5
Functionally Based Communicative Competencies 5
Field-Specific Writing Task Demands 6
The TOEFL Survey of Academic Writing Tasks 7
A Theoretical Perspective of Functional Communicative Competency 9
Perspectives from Contrastive Rhetoric 10

Design of the Writing Assessment Validation Study 11

Instrument Development 11
Administration Factors 13
Data Collection 13
Sample 13
Testing procedures 14
Scoring of the Instruments for the Direct Assessment of Writing 15
Scoring methods 15
Scorers 19
Scoring procedures 19
Psychometric and Interpretation Factors 20
Data Analyses 21

II. DEVELOPMENT OF RESEARCH INSTRUMENTS 23

Development of Instruments for the Direct Assessment of Writing 23

Development of Writing Tasks 23
Pretesting of Writing Tasks 27
Pilot Testing of the Eight Topics 29
Reading of Pilot Test Writing Samples 30
The training of readers 31
The scoring of papers 32
Conclusions and implications for the January reading 33
Final Selection of Topics 34
Formatting of the Test Booklet 34

Selection of An Indirect Measure of Writing Ability for the GRE Sample 35

Development of the Essay Reader Questionnaire 36

III. ADMINISTRATION OF EXPERIMENTAL TESTS 39

International Administration of Writing Samples 39

United States Administration of Direct and Indirect Measures of Writing Ability 41

Description of the Sample 41

IV. SCORING THE WRITING SAMPLES AS DIRECT MEASURES OF WRITING ABILITY 44

Preparation for the Essay Reading Weekend 44
- Planning for the Reading Weekend 44
- Sample Picking Sessions 45
- Chief Readers' Meeting 46
- Table Leaders' Meeting 46

The Essay Reading Weekend 47
- Holistic Scoring 48
- Discourse/Sentence Scoring 48

Cleanup Readings 49

Subject Matter Readings 49

Writer's Workbench Descriptive Scoring 50

Scoring of Other Instruments 51
- LSAT Indirect Measure 51
- Reader Questionnaires 51

V. RESULTS 53

Descriptions of Scores on the Conventional Tests 53
- TOEFL Scores 53
- GRE General Test Scores 53
- LSAT Writing Test Scores 55

Writing Sample Scores 55
- Means and Standard Deviations--Foreign Sample 55
- Means and Standard Deviations--English-Speaking U.S. Sample 55

Estimates of Score Reliability for Writing Samples 56
- Reliability of Holistic Scores 56
  - Interrater reliability 56
  - Reliability across topics 57
  - Reliability within language groups 58

Reliability of Discourse- and Sentence-Level (D/S) Scores 59
  Interrater reliability 59
  Reliability across score types and across topics 60
  Reliability within language groups 60
  Reliability across ESL and English readers 61

Reader Responses to Weekend Reading Questionnnaires 63

Correlations of Holistic Scores with Ratings of Subject Matter Experts 65

Exploratory and Confirmatory Factor Analyses 66
  Relationships of Writing Sample and TOEFL Mean Scores 67
  Relationship of Demographic Variables to Writing Sample and TOEFL Scores 68

Correlational Analyses 70
  Correlations with TOEFL Scores 70
  Correlations with GRE General Test Scores 71
  Writer's Workbench Analyses 73

VI. SUMMARY OF RESULTS AND CONCLUSIONS 76

Conclusions 80

Recommendations 81

Bibliography 83

Tables 90

Appendixes: 127

A. Writing Assessment Test Instructions and Topics

B. List of Readers for Reading Weekend
List of Subject Matter Readers

List of Tables

Table		Page
1	TOEFL Score Data for Total Sample of International Candidates and Three Language Groups	90
2	Scores on Writing Samples, TOEFL, GRE General Test, and LSAT Writing Test for Sample of GRE Candidates	91
3	Scores on Writing Samples, TOEFL, GRE General Test, and LSAT Writing Test for United States and International Samples of GRE Candidates	92
4	Scores on Writing Samples for Total Sample and International Language Groups	93
5	Criteria Used to Evaluate Written Assignments. Saturday and Sunday Reader Questionnaire Responses (in percentages of total of 50 respondents on Saturday, 51 respondents on Sunday)	94
6	Criteria Used to Evaluate Written Assignments. Saturday Reader Questionnaire Responses Prior to Reading Sessions (in percentages of total of 50 respondents and 24 ESL readers, 26 English readers)	96
7	Criteria Used to Evaluate Written Assignments. Saturday Reader Questionnaire Responses During Holistic Scoring (in percentages of total of 50 respondents and 24 ESL readers, 26 English readers)	98
8	Criteria Used to Evaluate Written Assignments. Sunday Reader Questionnaire Responses During Discourse/Sentence Scoring (in percentages of total of 51 respondents and 24 ESL readers, 27 English readers)	100
9	Reader Responses to Questions About Scoring Systems on Saturday and Sunday Questionnaires	102
10	ESL and English Reader Responses to Questions About Scoring Systems on Saturday Questionnaires	103
11	ESL and English Reader Responses to Questions About Scoring Systems on Sunday Questionnaires	104
12	Factor Loadings Obtained from the Principal Axes Factor Analysis. Seven Writing Sample and TOEFL Variables (N=560)	106
13	Factor Loadings Obtained from the Principal Axes Factor Analysis. Seven Writing Sample and TOEFL Variables. Arabic language group (N=139)	107
14	Factor Loadings Obtained from the Principal Axes Factor Analysis. Seven Writing Sample and TOEFL Variables. Chinese language group (N=230)	108

Table		Page
15	Factor Loadings Obtained from the Principal Axes Factor Analysis. Seven Writing Sample and TOEFL Variables. Spanish language group (N=191)	109
16	Correlations of Holistic Scores, D/S Scores, and TOEFL Scores (total sample of 542 candidates)	110
17	Means and Standard Deviations for Writing Sample and TOEFL Scores	111
18	Correlations of Demographic Variables with Holistic Scores, D/S Scores, and TOEFL Scores (total sample of 542 international candidates)	112
19	Correlations of Demographic Variables with Holistic Scores, D/S Scores, and TOEFL Scores (sample of 138 Arabic language candidates)	113
20	Correlations of Demographic Variables with Holistic Scores, D/S Scores, and TOEFL Scores (sample of 230 Chinese language candidates)	114
21	Correlations of Demographic Variables with Holistic Scores, D/S Scores, and TOEFL Scores (sample of 174 Spanish language candidates)	115
22	Correlations of Holistic Scores, D/S Scores, TOEFL Scores, LSAT Writing Scores, and GRE General Test scores (sample of GRE candidates)	116
23	Correlations of Holistic Scores Total and GRE Item Type Scores (sample of 132 cases)	117
24	GRE General Test Item Types Stepwise Regression Aanalysis for Holistic Score Total (N = 132)	118
25	Significant Correlations of TOEFL Section I (Listening Comprehension) with Writer's Workbench Variables, TOEFL Section Scores, and Writing Sample Scores for Four Writing Sample Topics	119
26	Significant Correlations of TOEFL Section II (Structure and Written Expression) with Writer's Workbench Variables, TOEFL Section Scores, and Writing Sample Scores for Four Writing Sample Topics	120
27	Significant Correlations of TOEFL Section III (Reading Comprehension) with Writer's Workbench Variables, TOEFL Section Scores, and Writing Sample Scores for Four Writing Sample Topics	121

Table		Page
28	Significant Correlations of Holistic Scores on Writing Samples with Writer's Workbench Variables, TOEFL Section Scores, and Writing Sample Scores for Four Writing Sample Topics	122
29	Significant Correlations of Discourse/Sentence Scores on Writing Samples with Writer's Workbench Variables, TOEFL Section Scores, and Writing Sample Scores for Four Writing Sample Topics	123
30	Writer's Workbench Stepwise Regression Analyses for TOEFL Section II. Structure and Written Expression	124
31	Writer's Workbench Stepwise Regression Analyses for Holistic Scores	125
32	Writer's Workbench Stepwise Regression Analyses for D/S Scores	126

I. INTRODUCTION

The ability to write clearly is an essential skill needed by undergraduate and graduate students. With the recognition that too many students pass through our educational system with only minimal English language competence, educators are reappraising their methods and redefining their objectives. Writing competence, in particular, is being addressed as a skill that is integral to effective communication. Therefore, researchers and educators recently have directed considerable effort toward the measurement of writing ability, and, in turn, to the understanding of its relationship to other cognitive skills. In the past, measurement of writing skills has been achieved largely by means of indirect measures--test items cast in the multiple-choice format. However, the definition of writing competence currently is being expanded and refined. Although multiple-choice measures provide some indicators of written language skills, they do so indirectly, in that students respond to writing tests by recognizing a correct answer among a finite set of alternatives. Because the act of writing involves the production of a written piece, actual writing samples, or direct measures of writing, now are viewed as a more appropriate means for assessing writing performance because they more nearly approximate real discourse.

Two major testing programs at Educational Testing Service (ETS), the Test of English as a Foreign Language (TOEFL), and the Graduate Record Examinations (GRE), provide scores on multiple-choice measures that contribute to decisions made during the postsecondary admission process. The purpose of this study was to determine the relationship of scores on direct and indirect measures of writing ability to scores on the TOEFL and the GRE General Test. This project is a response by ETS to the assessment concerns expressed by educators in the field--the examination of direct methods for evaluating writing skills and the relationship of these measures to other, more conventional measures of developed abilities.

The TOEFL was designed to assist an institution in determining whether a foreign applicant for whom English is a second language has attained sufficient proficiency in English to study at that institution, at either the undergraduate or graduate level. An important component of that general proficiency is the ability to communicate in written English. In the TOEFL examination, the Structure and Written Expression section (Section 2) is an indirect measure of writing ability. The GRE General Test, as one indicator of potential for graduate study, serves as an instrument for admission to graduate-level education for applicants who are either native or nonnative speakers of English. The GRE General Test provides scores that are intended to assess developed abilities in the verbal, quantitative, and analytical reasoning domains, but does not contain an indirect measure of writing ability as such. Thus the TOEFL and GRE General Test provide complementary functions with respect to the admission of foreign students to graduate programs.

A recent informal survey of professionals in the field of English as a Second Language (ESL) conducted by Hale and Hinofotis (1981) identified the measurement of productive skills (e.g., speaking and writing) as highly desirable for preadmission testing and placement decisions. A report by

Angelis (1982) reached the same general conclusion. Angelis surveyed graduate faculty in the two fields that enroll the largest number of foreign students, engineering and business. He found that graduate faculty in engineering listed writing highest on the list of student deficiencies. Business faculty also listed the writing deficiencies of foreign students as a major concern. Furthermore, many respondents believe that the TOEFL is not effective in providing information about productive skills such as writing, because the skills assessed in Section 2, which stresses knowledge of grammar and conventions, may not adequately reflect actual writing skills.

Two previous studies have investigated the relationship between TOEFL scores and actual essay performance (Pike, 1979; Pitcher & Ra, 1967). However, these studies do not satisfactorily address current concerns about, and conceptions of, the nature of writing and writing assessment. This earlier research correlated ratings of essays with subtests on an earlier form of the TOEFL examination (a pre-1976 version with five subscores); since Pike's study (1979), the Writing Ability and English Structure subtests have been combined on the basis of his data. Pike also concluded that the strong relationships found between essay ratings and the Writing Ability section of the test suggested that there was little need for replacing this section with a writing sample.

On the basis of the writing assessment research of Godshalk, Swineford, and Coffman (1966), which indicated that the ratings of essays vary from topic to topic, Pike used four different topics in his study. However, the writing demands of these topics might not be as appropriate to the objectives of the TOEFL today, when viewed from the standpoint of obtaining valid samples of functional writing competency. Two of the four topics, using pictures as stimuli, required sequential descriptions of events; the constraints of these topics are not likely to provide the student with sufficient opportunity to demonstrate writing ability. In providing a predetermined framework, these topics are not likely to tap the student's ability to organize ideas, a skill that also is not measured by the discrete objective items in the TOEFL; it thus follows that such topics would not be expected to supply information that would supplement the objective measures. The other two topics, the writing of a dialogue and the comparison of the advantages of city and country life, required the student to incorporate certain words in the writing sample, a constraint that seems to create unreasonable demands; the topics also may have imposed cultural demands that prevented some students from adequately demonstrating their writing skills. Finally, the time limits allotted to each sample (10 minutes) may have restricted the possibility of organizing cohesive discourse. Foreign students, in particular, should be allowed enough time to process their ideas from one language to another (Lay, 1982). Although the Pike study provided sound data to support the construct validity of the TOEFL, we now have acquired additional knowledge about the design of writing topics, knowledge that should enable us to structure topics to elicit performance of the skills the samples are intended to measure. Moreover, since Pike's study involved the relationships between performance on writing samples and the old form of TOEFL, these relationships should be reexamined for the new form.

The current project investigated the relationship of scores from a current TOEFL form and from the GRE General Test with scores on writing samples reflecting the kind of performance that would be required of beginning undergraduate and graduate students in the three fields enrolling the largest numbers of foreign students--business, engineering, and the social sciences.[1] The study builds on the information obtained in a previous TOEFL project, a survey of academic writing tasks (Bridgeman & Carlson, 1983) that investigated the kinds of writing skills required of students across different departments in United States and Canadian institutions of higher education. The results of the 1983 study are summarized in a subsequent section of this report that brings together other research findings in the realm of communicative competencies. An additional objective of the TOEFL writing survey was to design a study that would relate TOEFL scores to scores on student writing samples, using appropriate topics identified in the survey. The study focuses on nonnative students who take the TOEFL as part of the admission process for entrance into United States and Canadian institutions. A logical extension of this work includes GRE General Test scores for nonnative as well as native speakers of English. Contrasting the correlational patterns for nonnative speakers in various academic disciplines with those for native speakers within and across disciplines provides information that is useful for admission and placement of both native and nonnative speakers and for more meaningful interpretation and use of GRE General Test and TOEFL scores.

Foundations for the Design and Implementation of the Study--Research, Theory, and Practice

The primary purpose of this study was to determine the relationships of TOEFL and GRE General Test scores to the kinds of writing tasks that first-year students are expected to perform. These data provide important information regarding the construct validity of the GRE General Test and the TOEFL, information that should be useful to those who interpret the scores on these tests as well as to ETS test developers who may be considering the addition of direct measures of writing ability, in the case of the TOEFL, and of indirect and direct measures of writing ability for future GRE General Test forms. The study involved the collection of four writing samples from native and nonnative speakers of English who were seeking admission to undergraduate and graduate levels of education in the United States and Canada. In addition, recent GRE General Test and TOEFL scores were obtained for the appropriate groups of candidates (e.g., candidates for admission to undergraduate programs do not take the GRE). The TOEFL scores include an indirect measure of writing skill, the Structure and Written Expression section of the TOEFL; scores on a comparable indirect measure (a section of a retired form of the LSAT) were obtained for native-speaking GRE candidates. The standardized test scores then were related to holistic and analytic scores on the writing samples. The plans for the data collection procedures are described in the final section of this chapter, and the specific procedures that were implemented appear in subsequent chapters. Before the implementation of the study is presented in detail, however, the rationales for the design of this complex project are explained in this section.

---

[1] According to the 1980-81 survey conducted by the Institute of International Education (Boyan, 1981), approximately 26 percent of the foreign students in the United States are enrolled in engineering programs, 17 percent in business and management, and 8 percent in the social sciences.

The most significant and fundamental tasks for this research required (1) the design of writing assessment instruments and (2) the collection and scoring of writing samples with these instruments. Elaborate planning was necessary, since the validity and usefulness of the information gained by the data analyses would depend on the quality of the measurement process. To achieve the best and most appropriate assessment of writing skills, the study design took into account the numerous perspectives that the state of the art in the evaluation of writing ability has to offer. We combined the knowledge and experience accumulated by a variety of disciplines--writing assessment and instruction, psychological measurement, linguistics, contrastive rhetoric, and instruction in English as a second language (ESL). Each of these fields offers insights garnered from theory, research, and practice. Our first planning objective focused on the definition of competence in writing, a definition that emphasizes the situational context of writing assessment appropriate to the objectives of the TOEFL and the GRE General Test as indicators of a student's ability to write English. This definition was formulated on the basis of information drawn from the areas of writing assessment, communicative competency, and contrastive rhetoric. Our second planning objective required the design of a validation study that depended on the development of effective instruments to evaluate written competence and on rigorously implemented data collection and scoring procedures. The following section briefly summarizes the framework for formulating a functional definition of writing ability, including our survey of academic writing skills. The subsequent section describes the bases for the design of the validation study.

A Definition of Writing Competence

The term "measure" suggests the ability to assign a value, or number, to what is being evaluated. In any form of writing assessment, that measure is subject to error, since it is based on inferential judgments with respect to standards that define competent writing. The definition of what we are seeking to measure is achieved by circumscribing the characteristics of writing ability, given the limitations of the state of the art in the measurement of written responses produced by individuals. In writing assessment, experts in the field still are attempting to develop an objective definition of competent writing. It is important to recognize that competent writing is a construct, or concept, that requires careful definition in order to be measured. In addition, the definition of this construct may vary from instance to instance, in that competent writing is situational--it is defined by the specific task demands within the particular situation in which, and for which, writing ability is being assessed. When writing ability is evaluated, that ability itself is not measured directly, but rather, assessed on the basis of inferences drawn from an individual's performance.

As we sought to develop a working definition of writing competence in the context of the GRE General Test and TOEFL, we drew on six perspectives, as described in the following sections: the new paradigm for writing assessment, functionally based communicative competencies, field-specific writing task demands, the TOEFL survey of academic writing tasks, a theoretical perspective of functional communicative competency, and perspectives from contrastive rhetoric.

The New Paradigm for Writing Instruction and Assessment

One leader in the field of writing assessment, Odell (1981), recently redefined writing competence ". . . to mean the ability to discover what one wishes to say and to convey one's message through language, syntax, and content that are appropriate for one's audience and purpose" (p. 103). In the direct assessment of writing, the writer is presented with some form of written communication that designates the task(s) to be accomplished. This communication varies in the degree to which the specific demands of a particular task are described. Depending on the amount and kinds of information provided, the verbal statements to the writer also communicate expectations about performance; in turn, these statements reflect, in varying degrees, the standards or criteria that will be applied in the evaluation of the written product.

Because the characteristics that contribute to competent writing are situationally dependent, the elements of the writing task presented should be predicated on a definition of writing competence that is directly parallel to the specific objectives for evaluating writing within a specific situational context. These objectives and the context of the evaluation must be described and, subsequently, reflected by the design of the writing assessment measure. Since the present research was conducted under the auspices of testing programs that serve as preliminary indicators of a candidate's readiness to participate successfully in an English-based curriculum at the undergraduate and graduate levels of education, we sought to define writing competence from the standpoint of the objectives of these tests--the standpoint of functional communicative competency.

Functionally Based Communicative Competencies

Linguists who have investigated the dimensions of language teaching and testing (Canale, 1983; Canale & Swain, 1979; Munby, 1978; Walz, 1982) emphasize the approach of "functionally based communicative competency." Briefly defined, it entails the ability to use language to communicate effectively within the specific context in which the communication takes place; it is "functional," in that it "works," serving to convey what the person intended and resulting in appropriate receptive behavior (thought or action) by the recipient of the communication. This functional orientation provides an explanation for the observed discrepancies between knowledge of grammar and conventions and actual production on direct measures of writing skills.[1]

Field-Specific Writing Task Demands

Other researchers have focused their investigations concerned with functionally based academic writing task demands on field-specific requirements, with emphasis on English for specific purposes. For example, West and Byrd (1982) surveyed 25 engineering faculty members at the University of Florida to identify the kinds of writing assigned to graduate students during one academic year (1979-80). West (1982) also surveyed 33 engineering faculty members during the same year, asking them to rate American and foreign students on eight writing dimensions. These faculty ranked the performance of all foreign graduate students lower than the performance of American students on all the writing dimensions, except for quality of content. Making pairwise comparisons on the eight dimensions of foreign student writing, West ordered the dimensions from weakest to strongest as follows: (1) correctness of punctuation, (2) quality of sentence structure, (3) vocabulary size, (4) correctness of vocabulary usage, (5) quality of paragraph organization, (6) quality of overall paper organization, (7) quality of content, and (8) overall writing ability. We adapted these dimensions for use in our TOEFL survey of academic writing tasks (Bridgeman & Carlson, 1983), described in the next section.

In another study that typifies research in writing for academic purposes, Johns (1980) focused on the cohesive elements in written business discourse. Hill, Soppelsa, and West (1982), stressing the academic need for ESL students to learn to write experimental research papers, outlined an instructional approach that similarly aims at functional discourse. Pointing to the growing interest in English for specific purposes and in English for academic purposes, these researchers identified experimental research papers as important to academic and professional success in the sciences and social sciences. Another ESL instructional approach recently described by Spack and Sadow (1983) emphasizes the composing process and writing assignments that students will face in academic and professional situations.

---

[1] Recently a number of researchers have attempted to identify some of the writing tasks that are required of graduate and undergraduate students within functional contexts: Freedman (1979), Johns (1981), Kroll (1979), Ostler (1981), Weaver (1982). Their findings, which identified writing task demands within the contexts of their specific institutions, are described in a report of our previous research (Bridgeman & Carlson, 1983).

The TOEFL Survey of Academic Writing Tasks

The literature on functional communicative competency served as the basis for the design of a research project that would provide a definition of writing task demands in postsecondary academic settings. The primary objective of this project (Bridgeman and Carlson, 1983) was to identify and describe operationally the expectations of writing competence required of nonnative speakers of English at the beginning of their educational experiences in institutions of higher education in the United States and Canada. The information we gathered took into account the various factors that should be considered in defining communicative competence in writing--the functional task demands for which students are expected to be prepared, as well as the perceptions, sometimes culturally influenced, of those who evaluate them. Initially, informal interviews and the literature provided the basis for the design of a survey instrument that incorporated the full range of expectations of writing competence. The writing task demands, features of writing tasks (adapting West's dimensions of student writing), and types of writing sample topics were expressed in terminology that would communicate clearly to individuals in various disciplines. Subsequently, a representative sample of departments within institutions responded to the questionnaire, providing a basis for describing the domain of writing competencies expected of entering native and nonnative students.

The survey questionnaire was completed by faculty in 190 academic departments at 34 universities in the United States and Canada with high foreign student enrollments. At the graduate level, six academic disciplines with relatively high numbers of nonnative students were surveyed: business management (MBA), civil engineering, electrical engineering, psychology, chemistry, and computer science. Undergraduate English departments were chosen to document the skills needed by undergraduate students.

The major findings are summarized as follows:

- Although writing skill was rated as important to success in graduate training, it was consistently rated as even more important to success after graduation.

- Even disciplines with relatively light writing requirements (e.g., electrical engineering) reported that some writing is required of first-year students. Lab reports and brief article summaries are common writing assignments in engineering and the sciences. Longer research papers are commonly assigned to undergraduates and to graduate students in MBA, civil engineering, and psychology programs.

- Descriptive skills (e.g., describe apparatus, describe a procedure) are considered important in engineering, computer science, and psychology. In contrast, skill in arguing for a particular position is seen as very important for undergraduates, MBA students, and psychology majors, but of very limited importance in engineering, computer science, and chemistry.

- Faculty members reported that, in their evaluations of student writing, they rely more on discourse-level characteristics (e.g., organization of ideas, quality of content) than on word- or sentence-level characteristics (e.g., punctuation/spelling, sentence structure, vocabulary size).

- Discourse-level writing skills of natives and nonnatives are perceived as fairly similar, but significant differences between natives and nonnatives were reported for sentence- and word-level skills and for overall writing. A majority of departments reportedly use the same standards for evaluating the writing of native and nonnative students, although nearly a third of the departments reportedly use different standards.

- Respondents were asked to rate types of writing sample topics, to indicate their preference for topics that would most likely elicit evidence of the writing skills that would facilitate performance in academic contexts. (Two examples of each type were provided.) The 10 topic types represented a range of writing assignments: (A) personal essay, (B) sequential or chronological description, (C) spatial or functional description, (D) compare and contrast, (E) compare and contrast plus take a position, (F) extrapolation, (G) argumentation with audience designation, (H) describe and interpret a graph or chart, (I) summarize a passage, and (J) summarize a passage and analyze/assess the point of view. The clear favorite among the engineering and science departments was Topic H (describe and interpret a graph or chart). However, this topic was perceived as inappropriate by a majority of the undergraduate English faculty. Topic G (argumentation with audience designation) was the favorite among MBA programs; Type E (compare and contrast plus take a position) also was evaluated positively by the MBA programs and was the favorite among undergraduate English faculty.

- To obtain a summary picture of the relationships among topic types both within and between academic disciplines, the acceptability ratings were analyzed using a multidimensional scaling approach that accommodates differences between raters. Within each discipline, the pattern of responses to each topic type was compared to the pattern of responses for every other topic type. The positions of the topic types, as rated by the respondents, reflect the perceptions of the similarities and differences among the topic types. The mulitidimensional scaling suggested that the respondents reacted to the topic types as having two dimensions, one determined by the complexity of the task demanded by the topic type, and the other, by the degree of personal involvement required. Topic H can then be seen as a relatively simple and impersonal task. Topic E is a little above average on the complexity dimension and is a task requiring a relatively high degree of personal involvement in the topic.

In sum, the faculty members surveyed appeared to view student writing skills from the standpoint of functional communicative competencies. For example, the written products prepared by students in different disciplines may be considered competent to the extent that they meet the task demands--particularly kinds of writing assignments and certain skills—that are specific to a discipline. In addition, faculty members reported that written assignments were evaluated on the basis of discourse-level characteristics, rather than word- or sentence-level characteristics, and that they perceived the discourse-level writing skills of natives and nonnatives to be fairly similar. Grammatical competency, however, tends to influence evaluations of student writing to some extent, since respondents reported that nonnatives are more deficient in word- and sentence-level skills than are natives.

A Theoretical Perspective of Functional Communicative Competency

Our effort to define academic writing tasks required of entry-level students in postsecondary institutions also was based on the theoretical insights of Canale and Swain (1979; Canale, 1983). Canale proposes a framework that distinguishes three types of language proficiency: basic, communicative, and autonomous. He believes that the most fundamental problem in language assessment results from the lack of an adequate theoretical framework for language proficiency. He summarizes the recent work by Bruner and Cummins regarding language proficiency and poses a framework that builds on their work, with modifications. Cummins (1983) provides a revision and clarification that is more directly applicable to language proficiency, in that language tasks are classified into four primary groups: cognitively demanding/cognitively undemanding and context-embedded/context-reduced. The context continuum for the classification of tasks ranges from context-embedded, which involves a "shared reality" or common world knowledge, to context-reduced tasks that ". . . require greater reliance on linguistic cues to meaning and on the propositional and logical structure of the information involved rather than on shared (or even existing) reality" (p. 337). The cognitive continuum ranges from tasks demanding little active cognitive involvement to tasks demanding much active, complex cognitive processing. This representation of language tasks clearly resembles the two-dimensional representation of writing topic types as perceived by academic respondents to our TOEFL survey of academic writing skills, in which the ordering of the ratings of topic types suggest two dimensions--cognitive complexity and personal involvement.

The perspective of functional communicative competency, in combination with other theoretical insights and research findings reported in this chapter, contributed several propositions that were the basis for design of this writing assessment research. The propositions are the following:

o Performance, which can be assessed in various ways, serves as an indirect means for evaluating language proficiency. The kind and degree of language proficiency being measured by a specific task are determined by the nature of that task.

- To evaluate performance on a task, the dimensions of that task, which condition the performance elicited, must be specified clearly.

- The following elements of a task that will be used to infer kinds and degrees of language proficiency must be accurately described, to the extent possible, both to the individual whose performance is being assessed and to the individual(s) who will evaluate that performance:

  The nature of the task demands, in terms of cognitive complexity and degree of personal involvement required.

  The nature of the linguistic performance that is expected to be elicited by the specific task, with the reservation that the linguistic performance that will be observed is what the examinee has produced within a specific context. That linguistic performance cannot necessarily be generalized to an evaluation of overall linguistic performance in (in this case) the written mode of communication.

  The nature of the hypothesized communicative situation in which the task places the examinee; e.g., the stated or implied purpose and audience to be addressed.

  The nature of the testing situation and all aspects of that social context that might influence differentially performance on the task; e.g., time limitations that do not allow for full organization and revision, the score on the task as one determinant of admission to an institution.

  The methods and procedures used for assigning scores to performance, which provide reasonable restrictions on score interpretation. The scoring method, for example, should reflect the scorers' appreciation of the dimensions of the task, the task demands, and the specific performance features that can be validly evaluated.

Canale (1983) proposes that the general framework originally posed by Canale and Swain (1979) with reference to communicative language also would be useful to other approaches to language. This distinction between communicative competence and performance is essential to the evaluation of language proficiency.

Perspectives from Contrastive Rhetoric

Another area of research that has explored the academic task demands required of nonnative speakers of English has been termed "contrastive rhetoric." In this area, rhetorical patterns across cultures are identified and compared (Kaplan, 1972, 1976, 1977, 1982). The results of studies of contrastive rhetoric provide somewhat mixed evidence, some rejecting and others supporting the underlying assumption that the structural differences between the native language and the foreign language may interfere with the learning of the foreign language. We reviewed several representative papers in this area in order to take cultural differences into account.

The work of Buckingham (1979), Lindstrom (1981), Pearson (1981), Takala, Purves, and Buckmaster (1982), and Purves (1984) particularly informed the process of topic selection and training of readers. The perspectives of cultural relativity provide a framework that influenced our decisions regarding the design of writing assessment tasks, the scoring of the collected writing samples, and the interpretation of results. Cultural differences in response to the demands of a writing task were taken into account as we attempted to identify and control the various parameters influencing the assessment of the writing performance of students from different international cultures. These parameters are described in the following section.

Design of the Writing Assessment Validation Study

The design of a writing assessment program is influenced by practical considerations such as costs and staffing; whatever the limitations imposed for the sake of efficiency, the interpretation of the results of any writing assessment must be conditioned by the factors that may have contributed to the results. Some of these parameters of a writing assessment program can be controlled, or accounted for, by good advance planning; others that cannot be controlled should at least be recognized as exerting possible effects on the outcomes of the assessment. The design of an investigation based on samples of writing ability requires the implementation of carefully planned procedures. As proposed, this validation study was executed in a series of stages: instrument development, administration of experimental tests, scoring of direct assessment instruments, scoring of other instruments, and analyses of data. These stages, summarized in subsequent chapters of this report, are briefly described here.

Instrument Development

As we became involved in the development of measures for the direct assessment of writing performance, we considered the additional perspectives afforded by practice, research, and current theory regarding design of writing prompts. A considerable amount of literature is devoted to the design of writing test prompts, as summarized by Ruth (1982). At ETS, another source of knowledge for this study was the experience of practitioners who have conducted large-scale writing assessment programs. This expertise represents the state of the art in the design of writing assessment tasks; however, much research remains to be conducted regarding to what extent the parameters of a writing assessment instrument influence writing performance.

The writing stimulus, or the verbal statement that elicits the specific writing performance being targeted, requires careful development and pretesting. Pretesting of the writing stimulus is essential--topics may appear superficially to achieve the desired results, but actual writing samples obtained from a representative population of students may yield surprising information about how the topic is perceived and the nature of the responses that are produced. Pretesting in this instance influenced our judgments about how well the following objectives were being met:

- The mode of discourse or type of writing assignment that the task presents (e.g., personal essay, persuasive argument) should reflect the expectations of writing required in undergraduate and graduate academic work in the United States.
- The writing tasks should avoid content with cultural bias, culture-bound vocabulary and concepts that might penalize a nonnative speaker, as well as topics that evoke heavily emotion-laden responses.
- The statement of the task should clearly communicate the expectations of writing performance demanded by the writing stimuli.
- The expectations for writing performance should be reasonable, given time constraints.
- Students would be asked to write on all four topics to elicit equivalent, comparable performances and to avoid eliciting differential performance within modes in response to different topics, which would invalidate the assessment.[2]

Our survey of academic writing tasks provided the basis for the development of writing assessment instruments. The survey enabled us to define writing competence functionally in terms of the writing tasks that beginning postsecondary students would be expected to perform and the measurement objectives of the TOEFL and GRE General Test. In addition, the survey guided us in the selection and implementation of the parameters influencing the measurement of writing skills, such as specific approaches to scoring, that were critical to this writing assessment data collection.

The survey indicated that no single essay topic type was universally accepted by all the academic disciplines surveyed. In the multidimensional scaling, Types H and E were further apart in the space than any other pair of types, suggesting that they were perceived as distinctly different tasks. Thus the Type E topic type was selected to serve as an effective contrast to the Type H topic type; since departments perceived these two types as distinctly different, it seemed likely that writing samples elicited by Types H and E elicited different writing skills, as well.

---

[2] A full discussion of these task factors will appear in the chapter, "Testing ESL Student Writers" (Carlson and Bridgeman, in press).

For this project, we proposed to develop two topics of Type H and two topics of Type E to which each student would respond. To administer topics that would most effectively meet the measurement objectives of the study, several topics of each type were developed and pretested. The pretesting allowed us to judge the topics in relation to the criteria discussed in this chapter.

Administration Factors

Major factors in test administration that contribute to the outcomes of writing assessment were taken into consideration:

- The physical layout of the writing stimulus was designed to give writers the opportunity for prewriting tasks of planning and organization and to suggest the expected length of the writing sample.
- Directions to adminstrators were designed to minimize such adverse conditions in the testing room as uncomfortable temperature, poor lighting, noise, and poor writing surface.

These factors are critical to any testing situation but assume greater importance when students are asked to generate and produce written responses.

Data Collection

Most of the data collection procedures that we had proposed to carry out were accomplished, with the exception of a few practical modifications.

Sample

We obtained a total sample of candidates for undergraduate and graduate study representing three language groups (Spanish, Arabic, and Chinese[3]) plus a group of native-English-speaking graduate students from the United

---

[3] Most Chinese TOEFL candidates are from Taiwan, but few undergraduates are tested in Taiwan. Large numbers of Chinese candidates for admission as undergraduates come from Hong Kong. Thus, we anticipated that Chinese graduate candidates would be drawn from test sites in Taiwan, and undergraduates from Hong Kong. This would provide for the greatest generalizability of the results to the actual TOEFL population. However, given the known differences between education in Taiwan and Hong Kong, the confounding of location with undergraduate status must be considered when the results are interpreted.

States. The group of students applying for admission at the graduate level was to be further subdivided into three major field categories: business, "hard" science, and social sciences/humanities. Some subsamples presumably are of greater interest to the TOEFL program, while other subsamples are of more interest to the GRE program. The sample that was obtained, however, contained few business majors; these candidates were included in the social science/humanities classification, resulting in two major field categories.

We proposed to test samples of approximately 270 students from each language group, 70-75 candidates for admission as undergraduates and 200-210 candidates for admission as graduate students. More students were needed in the graduate category because this group would be divided and analyzed separately according to academic majors; undergraduate students would be treated as a single group for analysis purposes. As described in the chapter on data collection, the actual total sample (662) obtained was smaller than anticipated; however, the sizes of the total sample and native language subgroups were sufficient for the statistical analyses. The sample sizes varied, depending on the amount of missing and complete data, for each of the several analyses. The detailed descriptions of the resulting data collection and analyses are reported in subsequent chapters.

The GRE/TOEFL group included candidates for admission as graduate students who had taken (or planned to take) both the TOEFL and GRE examinations. The TOEFL-only group included foreign candidates for admission to institutions in the United States as undergraduate or graduate students. The GRE-only group included native-English-speaking candidates who were candidates for graduate admissions to institutions in the United States. GRE scores were obtained for native-English-speaking candidates, whereas both TOEFL and GRE scores were obtained for candidates for admissions in both the domestic and foreign samples who had taken the TOEFL as well as the GRE. Native students responded to the writing assessment instruments at universities in the United States; nonnative students responded on the day on which they took the TOEFL at international test centers.

Testing procedures

Because language skills can change dramatically in a relatively short period of time, testing students in the United States some months after they took the TOEFL in their native countries might lead to inexplicable confounding and uninterpretable results. Instead, we tested students at foreign centers as close in time as possible to when they took the TOEFL. GRE scores should be less subject to short-term fluctuations, and any student who had taken the GRE up to six months before the TOEFL or who was scheduled to take the GRE up to six months after the TOEFL was eligible for inclusion in the sample.

The international centers were selected, with the assistance of program staff, based on the following criteria: having candidates from the desired language groups, having candidates representing diverse ability levels, having a reasonable balance of undergraduate and graduate candidates, and having substantial numbers of GRE (recent past or potential) candidates. The procedures for selecting and inviting the candidates varied, depending on the specific conditions at each test site, as described in Chapter III.

Writing samples from GRE candidates in the domestic sample were collected during special testing sessions at five major university testing centers after we had identified and selected recent GRE General Test takers. Since the GRE General Test does not contain an indirect measure of writing skills, the GRE candidates at domestic sites also took a brief objective test of writing skills, a retired form of a test of writing skills formerly used by the Law School Admission Testing program. Thus we were able to compare indirect measures with direct measures of writing for the native GRE candidates, as well as for the nonnative TOEFL and GRE candidates.

As proposed, each native and nonnative English-speaking candidate produced four writing samples, two samples per topic type. We collected this number of samples in order to elicit a reasonable representation of writing skills, as well as an indication of the degree of consistency in the performance of individuals across similar and different tasks. We recognized that the samples ideally should be obtained at more than one sitting, to avoid fatigue and uniform responses that students might show because tasks are consecutive (Diederich, French, & Carlton, 1961; Godshalk et al., 1966). The one-day testing situation was unavoidable, however, because of the logistics (and subsequent attrition) involved in asking students at the international testing centers to return on another day. Thus the writing sample topics were designed to be sufficiently different to discourage mechanical responding. The distinctly different topic types also were expected to elicit different writing skills, particularly since the task requirements were to be carefully phrased to emphasize their different expectations.

Scoring of the Instruments for the Direct Assessment of Writing

Scoring methods

Selection of an appropriate scoring method for a writing sample depends on the purposes of the assessment. A holistic evaluation (i.e., a single score representing the overall impression created by the sample) may be more efficient for making selection or placement decisions, whereas a more analytic framework (i.e., separate scores for a number of different organizational and grammatical features of the sample) may be more useful for providing diagnostic information to teachers. Although other methods (e.g., error counts) may yield more objective scores as a rough index of second language proficiency, they may be poor indicators of functional communicative competence.

Holistic scoring is impressionistic, but it is not haphazard. Considerable care must go into selecting sample essays (range finders) that represent each point on the score scale, and thorough training of the readers is necessary. Such training involves discussion among the readers to reach consensus on the criteria. During a reading session, continual checks must be made to ensure that no reader is straying from the standards originally set. Since the scorer judgments are subjective, each essay should receive at least two independent readings. The scores from the two readers are typically added together to form the single holistic score.

Holistic evaluations may be influenced by a number of features of an essay, including content, organization, sentence structure, and mechanics. A study by Freedman (1979), in which essays were rewritten so that they exhibited strengths or weaknesses on each of the preceding four traits, indicated that content and organization had the greatest influence on holistic scores. Mechanics and sentence structure influenced scores only if the essay was well organized. However, generalizing from studies based on essays written by native speakers to essays written by ESL students may be unwarranted. Breland and Jones (1982) used a set of 20 scores classified as discourse, syntactic, or lexicographic characteristics to predict holistic scores that had been independently assigned. Paralleling the findings of Freedman, they found that the discourse characteristics were the best predictors of holistic scores. However, unlike Freedman, Breland and Jones included a group of essays written by Hispanic ESL students. In this group, syntactic and lexicographic scores were relatively much more important. Subject-verb agreement and range of vocabulary were particularly strong correlates of holistic scores in the Hispanic group. This finding may simply reflect the greater range of syntactic and lexicographic skill found in an ESL population. Regardless of the reason for the differences in the ESL group, this study serves as a useful reminder that even well-established "facts" concerning the scoring of writing samples may have to be modified for ESL populations.

In native speakers, for example, organizational skills usually parallel mechanical skills, and it is unusual to find highly organized essays written by students with very poor grammatical skills. With students for whom English is a second language, a greater disparity between organizational skills and mechanical competence in English would not be unreasonable to expect.

If a single holistic score is to be used, the raters must agree on how to score essays that present a large discrepancy between organizational and mechanical skill. They must also agree on which mechanical errors are most serious. This judgment of error gravity may stem from a strictly functional communication point of view (Does this error interfere with what the author is trying to say?), or it also may penalize errors that are stylistically undesirable (e.g., redundancy, run-on-sentences). In addition, raters must agree on how to evaluate essays that contain complex sentence structures, and in which the writers make errors in trying to write complex sentences, versus essays that use only simple sentences but contain few errors. In her research, Greenberg (1983) noted that ability to avoid errors predicted teachers' quality ratings better than the writer's ability to handle complex syntactic structures. She found that one major problem consisted of word form errors. Shaughnessy (1977), in fact, recognized that word form errors exemplify "advanced errors." Such errors indicate attempts to acquire formal academic vocabulary in spite of the risk of making errors. Thus more competent writers may commit more errors, yet may be penalized by raters who focus on the lack of errors as a predominant feature of good writing. During the training for holistic scoring, discussion about errors should be limited so as not to interfere with the process of reading for total impression, and to ensure that particular features of writing do not unduly influence that total impression.

Despite the most rigorous procedures in the training of scorers, holistic scoring schemes inevitably require some degree of subjective judgment, and these subjective judgments may be particularly difficult when the writer and reader (scorer) do not share a common set of cultural conventions and expectations. These conventions go far beyond mere differences in grammatical rules. The work of Kaplan (1966) clearly demonstrated cultural differences in patterns of logic used to order ideas within paragraphs. For example, Kaplan suggests that Anglo-European expository essays typically follow a linear development. In contrast, paragraph development in Semitic languages is based on a complex series of parallel constructions of coordinate rather than subordinate clauses. Oriental essays use an indirect approach; the reader is told how things are not, rather than how they are. In French and Spanish essays, Kaplan noted more digression and introduction of extraneous material than would be considered acceptable in an English essay. Thompson-Panos and Thomas-Ruzic (1983) recently noted certain contrasting features of English and written Arabic that may contribute to perceived weaknesses in the writing of Arab ESL students. For example, paragraph development in Arabic languages consists of a series of parallel constructions connected by coordinating conjunctions, thus deemphasizing the use of subordination that is valued in English paragraph organization.

ESL teachers who are aware of distinct cultural patterns may assign essay ratings that differ significantly from ratings of English teachers with no ESL experience. On the other hand, if the criterion for competence is success in a standard course in a United States university, the "insensitive" ratings may better predict academic performance than the culturally sensitive ratings. In this study, we compared ratings by ESL readers with ratings by readers whose predominant experience is with native speakers of English. In addition, these ratings were compared to ratings given by faculty members in engineering and the social sciences. The classic research of Diederich et al., (1961) suggests that, even among native speakers, different "schools of thought" exist among readers, and that certain professions are more likely to emphasize a particular characteristic. For example, lawyers appear to focus more on organization, whereas editors tend to focus on style and wording. In our research, the essay readers completed a questionnaire intended to identify the features they attend to when evaluating a composition.

Because analytic scoring yields more scores than holistic scoring, it is potentially more valuable for prescribing educational interventions for individual students. One scoring scheme that has been used extensively with ESL students provides separate scores for content, organization, vocabulary, language usage, and mechanics (Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey, 1981). Other analytic scoring schemes provide for even finer-grained analysis. However, the apparent advantage of several separate scores is frequently an illusion; the reader's general impression is likely to influence ratings on each of the "separate" aspects being evaluated. In addition, analytic ratings are very time consuming. Wiseman (1949) found that four general impression markings were equivalent in time and effort to one analytic marking. As noted previously, despite considerations of efficiency, a single holistic score may not adequately describe an ESL student with discrepant organizational and mechanical skills. Further research is needed to determine the best compromise between a single score and a complex analytic scoring scheme, as well as which kinds of scores are more appropriate to specific situational contexts.

The most promising means for the objective scoring of essays may be by computer software such as Bell Labs' "Writer's Workbench" (Cherry, Fox, Frase, Gingrich, Keenan, & Macdonald, 1983; Kiefer, & Smith., 1983). This sophisticated word processing tool can identify such features as spelling errors, overuse of a particular word, and sentences that are consistently too long or too short. Analysis of these structural features might help some writers improve their writing. However, this kind of computer program cannot judge how well a piece of writing accomplishes its main purpose of communicating with its intended audience, nor can it evaluate features such as development and organization. The subjective impression of coherence that a reader "receives" from the written communication cannot be duplicated by a mechanical count of cohesive elements (Carrell, 1982).

In this study, essays were scored holistically using scoring techniques developed at ETS (Godshalk et al., 1966) and refined over the years as a standard procedure in several ETS testing programs. In holistic scoring, judgments are made about qualities of the essay as a whole rather than by obtaining numerical counts of specific features. But holistic scoring does not imply that only one global score may be assigned to each essay. Several different characteristics of an essay may be evaluated holistically. The faculty members responding to our survey of academic writing skills indicated that most departments would like to see more than one score assigned to each essay. Therefore, we also planned to generate three holistic scores for each essay: one for content and quality of ideas, one for grammatical and mechanical errors, and one for organization and coherence. Subsequent to the proposal, Breland and Jones' (1982) research, alluded to in a previous section of this chapter, suggested that no more than two scores could be assigned independently to writing samples, as did consultation with other experts in holistic scoring procedures. The scoring procedures we adopted are discussed in the chapter devoted to scoring the direct measures.

In addition to the holistic scores, we also planned to obtain simple analytic scores (e.g., total essay length, average sentence length, subject-verb agreement). However, instead of obtaining analytic scores for papers using human judges, a representative subsample of papers written in response to each topic were analytically analyzed on Bell Laboratories' Writer's Workbench software at Colorado State University. Joy Reid, an ESL faculty member and researcher, supervised these analyses, while Roberta Scott, a composition instructor, keyed the papers into the computer. A complete description of this procedure appears in a subsequent chapter.

Scorers

The scorers included individuals experienced with assessment in ESL and English composition, including a core of scorers experienced in holistic scoring.

In order to obtain additional and independent scores for the writing samples, we also obtained ratings for a subsample of papers from faculty members from the two academic disciplines with the largest foreign student enrollments. They were asked to evaluate the papers from the perspective of writing skills exhibited at the time of admission to their program, rather than from the perspective of writing skills expected to be developed as students develop discipline-specific writing skills. These ratings by faculty members made it possible to compare scores assigned by subject matter experts with scores assigned by writing experts, providing some indication of the extent to which points of view regarding writing competence reflect different perspectives within these disciplines.

A slight change from proposed procedures was made with regard to the rating of papers by subject matter readers because of the dearth of potential business majors. Instead of using subject matter experts representing business and the hard sciences, four faculty members in each of two disciplines, the social sciences and hard sciences, assigned ratings to representative samples of papers written in response to two essay topics, one of each of the two types.

Scoring procedures

The holistic scoring procedures basically were those outlined in the proposal. However, some alterations were made in order to ensure the quality of the readings.

Because the TOEFL program is considering the addition of a writing sample to their operational testing program, the objectives of the holistic scoring were twofold: (1) to obtain valid and reliable scores to contribute to the statistical analyses relevant to the research objectives regarding the construct validation of GRE and TOEFL scores, and (2) secondarily, to provide information about scoring procedures that would be useful to an operational writing assessment program. The holistic scoring sessions were carefully designed in the light of these objectives. The basic operational procedures used at ETS were employed, but additional complexities were introduced because a controlled research design also was imposed.

Psychometric and Interpretation Factors

Although psychometric considerations of reliability and validity are essentially the same for ESL essays as for essays written by native speakers, the unique cultural and linguistic characteristics of ESL students require special attention.

Reliability or consistency of essay scores can be assessed in a number of different ways (intrarater, interrater, across topics within genre, across genre). Intrarater reliability indicates how consistent a single rater is in scoring the same set of essays twice with a specified time interval between the first and second scoring. Interrater reliability estimates the extent to which two or more raters agree on the score that should be assigned to an essay. When essay writers and raters represent different cultural perspectives, interrater reliability is likely to be lower than when both essay writers and raters come from a homogeneous group. But even if interrater reliability is perfect, the claim cannot be made that the essay test is perfectly reliable. Other factors such as variations over time, from one topic to another, and from one sample of students to another also must be considered.

Intertopic reliability assesses the extent to which the rank ordering of student scores depends on the topic. Scores will vary from one topic to another even within the same general topic type (e.g., compare and contrast). A relatively small intertopic variation in a group representing a single cultural group may become quite pronounced in a culturally diverse sample if one of the topics is particularly evocative for students from one culture. For example, a topic comparing life in a democracy to life in a dictatorship may represent an abstract academic exercise for North American students but may stimulate an intense personal reaction from students from Central America. In addition, variations from one topic type to another (e.g., narrative vs. persuasive) may be even more influenced by cultural factors.

High reliability does not provide sufficient evidence that a test is valid. Instead, the test may be measuring consistently a variable that is not the criterion of primary interest. Thus, a 30-minute writing sample might be judged reliable, but it might not serve as a valid indicator of the student's ability to write a long paper without limitations on time and with an opportunity to make extensive initial drafts.

As Cronbach (1971) has noted, it is not tests that are validated but rather interpretations of data from tests used in specific contexts. Scores from an essay test may be valid for one purpose but not another. For instance, a test that serves as a valid indicator of skill in writing a narrative essay may have little value in predicting a student's ability to meet the writing demands in a graduate engineering program. Furthermore, a test that is considered a valid predictor of success in meeting the writing demands of undergraduate study for native speakers may or may not predict with comparable validity for ESL students.

Optimally, validity should be determined by establishing that a test is measuring the same performance objective that a good external criterion also is measuring. When the parameters that condition a measure of writing skills are taken into account, the external appearance of a writing sample topic, or its face validity, is not sufficient to ensure the validity of the performance that is intended to be measured. An objective means for determining the validity of scores on a writing sample can be achieved by correlating these scores with scores on other measures that have been demonstrated to predict well to the same criterion. This criterion, likewise, must have evidenced validity and reliability. One frequently used criterion of academic success, such as the grade point average, may not meet consistently the constraints of validity and reliability. Instead, valid and reliable scores on an established test that has been shown to predict to the criterion (i.e., grades) may serve as a more objective indicator for validating writing sample scores. The validity of scores for writing samples that are included in standardized tests, for example, is established by demonstrating that the scores are highly correlated with scores on indirect measures of writing ability.

Ideally, however, scores on direct and indirect measures would not be perfectly correlated. Because a writing sample requires the production of a composition in contrast to the recognition of correct responses on a multiple-choice test of writing ability, we would not expect the two types of test to assess identical skills. Instead, they would be highly correlated because some of the skills they are measuring overlap and reflect a form of "general" writing ability. In addition, writing samples would be expected to contribute additional information about writing performance that is not yielded by an objective test, thus explaining an imperfect correlation.

Test validation is a process of accumulating evidence to support inferences made from test scores, reflecting the value of a test for an intended purpose; more sources of evidence are better than fewer. For this study, we intentionally planned to score the writing samples in various ways, and to relate these scores to other measures, in order to obtain as much information as possible regarding the validity of direct measures of writing in the TOEFL and GRE contexts. The different procedures and analyses are discussed in subsequent chapters.

Data Analyses

We performed several statistical analyses of the data, consisting of correlational and factor analyses. The specific analyses and results appear in the final chapters of this report. The data analyses were conducted to reveal the degree of relationship among several variables--GRE section scores and item type scores, TOEFL total and section scores, scores on indirect measures of writing ability (included in the TOEFL and as a separate test for native speakers of English), and the different scores derived from direct measures of writing ability. In addition, the obtained relationships were examined with respect to the different language groups (Arabic, Chinese, Spanish, and English). The objective of the data analyses was to provide information about the content and construct validity of the GRE and TOEFL examinations; in particular, the data would suggest the extent to which writing ability contributes to GRE and TOEFL test scores.

II. DEVELOPMENT OF RESEARCH INSTRUMENTS

In preparation for making comparisons of direct measures of writing ability with indirect measures and with TOEFL and GRE scores, writing tasks were carefully designed for the assessment of performance on writing samples. The procedures used in designing, pretesting, and pilot testing the writing tasks are described in the following section of this report. For the sample of students for whom English is the primary language, and for whom, therefore, only GRE scores were available, a multiple-choice section retired from the Law School Admission Test was used to provide the indirect measure of writing ability; that test is described in the second section of this chapter. For the readers of the student papers, who represented two different disciplines--ESL and English composition--a questionnaire was developed to survey the readers' general perspectives on the evaluation of writing and their reactions to scoring the writing samples in this study; this questionnaire is described in the final section of this chapter.

Development of Instruments for the Direct Assessment of Writing

This process, a critical element of the study, demanded attention to, and, to the extent possible, control of the numerous factors that influence a direct assessment of writing ability. Besides heeding the many considerations that normally influence the design of the writing task, we needed, through pretesting and pilot testing of topics, to test our assumptions concerning the writing performance that would actually be elicited by our particular topics and the tasks they presented. The topics then were pretested, resulting in the selection of a reduced number of topics with the potential to tap writing performance effectively. Furthermore, these topics were pilot tested, and the resulting writing samples were scored in an essay reading that focused on the writing performance elicited by the topics. Eventually, the final topics that were selected for administration to the large sample of international and U.S. students were refined and formatted in carefully designed test booklets. The detailed descriptions of these procedures are presented in this section of the report.

Development of Writing Tasks

This effort was based on the information obtained in the survey of academic writing tasks summarized in the preceding section. Although our survey indicated that no single type of writing sample topic was universally accepted by all academic disciplines surveyed in the study, two topic types were selected as most representative of the kinds of writing tasks that would be useful performance indicators for institutions of higher education during the admissions process. As described previously, the compare/contrast topic type was selected as an effective contrast to the graph/chart topic type; the fact that departments perceived these two types of tasks as distinctly different suggests that writing samples elicited by these tasks may demonstrate different writing skills as well.

Several parameters affecting the design of a writing task were taken into acccount while the staff wrote the preliminary set of topics:

- The content of the topics needed to be equally accessible to the variety of students who would be responding to it. Since the nonnative speakers of English in the research sample would come from different cultural backgrounds, the content implied by the topic could not favor a particular set of personal or cultural experiences. Subtle biases in the topics were avoided by eliminating topics that suggested controversial social norms (e.g., family size reflecting family planning), or social conventions (e.g., the Dewey Decimal System in a library), or cultural perspectives (e.g., assumptions of American middle class views).

- In addition, the terminology in the writing tasks was to be free of vocabulary and concepts that required specialized knowledge for an effective response to the topic. Because the primary objective of the topic was to stimulate performance representative of the student's writing ability, culture-bound terms and concepts present artificial obstacles to that performance. Similarly, writing tasks that posed a high level of reading difficulty and vocabulary were revised or eliminated, since reading ability and differential standards of English vocabulary mastery would confound the assessment of writing skills.

- Topics also were designed to diminish the possibility that emotional responses would be evoked by the subject matter or by the hidden agenda of the task. Topics stimulating a highly personal reaction could either create an emotional obstacle deleterious to performance or lead to the production of a writing sample in the form of a personal essay rather than in the mode of discourse that was intended.

- The subjects of the tasks needed to be sufficiently compelling to the writers and, eventually, to the readers of the writing samples. Each subject was chosen to be interesting enough to engage the writer's interest and promote some latitude in responding—providing the writer with a relatively challenging task and the reader with a range of performance to be evaluated.

- From the standpoint of the evaluation of writing samples, experience with large-scale testing programs at ETS indicates that the most effective prompt for the writing task is one that elicits an optimal range of responses. Ideally the range of responses should be sufficiently broad to make distinctions, yet not so broad that the responses are too divergent to compare on an evaluative scale or so narrow that the writers are limited in demonstrating their abilities to deal with the assignment. Although a writing task may have the appearance of meeting this requirement, its success can be verified only by collecting a representative number of writing samples and observing the actual performance of writers who respond to the task.

- The length and specificity appropriate to the writing should be conveyed to the writer by the topic and accompanying cues. The topic should communicate enough about expectations to elicit the writing performance that is desirable for evaluation. If the task fails to communicate these expectations to the writer, readers will have difficulty accommodating their judgments about the writing samples that are obtained to their evaluation criteria, and writers may be penalized inadvertently for failing to address the task appropriately.

- The mode of discourse or form the written product will take (e.g., personal essay, persuasive argument) also is conveyed by the writing stimulus. This constraint on the writing task is determined by the objectives for evaluating writing performance. For example, if the ability to develop a personal essay is to be an important objective of a writing assessment, the task should be structured to elicit personal writing. However, in this instance the previous survey of writing tasks indicated that specific types of writing are valued within specific academic contexts. Since the goal of this research was to obtain writing samples that elicited these forms of writing, the stimulus needed to be designed so that student writers would respond with the expected forms. The form the written product takes is conveyed not only by the explicit instructions to "summarize" or "describe" but also by the content of the topic that serves as the vehicle for expressing ideas within format. Thus the writing stimuli need to be written with consideration for the kinds of writing that might be anticipated when prompted by a particular ideational structure. For example, a compare/contrast topic might be so emotionally charged that most students would produce a personal essay in response. Although most writing tasks actually consist of a combination of modes of discourse, the dominant mode that is to be evaluated must be emphasized.

- Many experts in English composition stress that the purpose and audience for the writing sample should be specified. Because we were attempting to attend to the potential cross-cultural differences of the students who would be writing for this study, the audience and purpose were not stated specifically. The designation of a specific audience and purpose may have introduced cultural and experiential bias (e.g., liberal arts candidates vs. engineering candidates)--audience specification must be explicit, but also appropriate. In addition, we made the assumption that the TOEFL and GRE General Test candidates are keenly aware of the purpose and audience for a writing task that is a part of an examination taken for the purpose of demonstrating a level of

English proficiency for admission to institutions of higher education in the United States. Eventually, as we critically analyzed writing samples obtained in the pilot testing, we determined that audience and purpose needed to be more clearly prescribed for the chart/graph topics in order to clarify the expectations for this topic; students had responded differently to the original prompt--some wrote a descriptive piece, whereas others presented interpretations of the data. Refinement of this topic resulted in writing samples that more nearly met the task demands in the final administration. Our experiences with topic design as a result of pretesting and pilot testing are described in more detail in a subsequent section.

- From the measurement standpoint, one writing sample is equivalent to a one-item test. In interpreting the results of an assessment, serious validity and reliability concerns restrict generalization from such a limited sample of performance. Ideally, the assessment of writing ability should consist of more than one item.

  Furthermore, before decisions can be made on the basis of this performance, we need to be assured that the sample that has been obtained within the constraints of a testing situation is representative of the individual's writing ability.

  Similar measurement concerns are raised if comparisons are made among scores for students who have taken tests composed of different questions. For multiple-choice tests containing multiple items, this equating problem can be resolved statistically; therefore, for example, a score on one form of the TOEFL or GRE examination is directly comparable to a score on another test form. In large-scale testing programs at ETS, scores on writing samples are equated through the multiple-choice test. For scores on different writing samples alone, however, the psychometric capability for equating items has not been developed.

  When students are given the opportunity to write in response to different tasks, we cannot be assured that each student has been given equal opportunity to demonstrate writing despite the apparent comparability of assignments. Numerous uncontrolled variables are introduced in this instance, such as the differential effects of topics, modes of discourse, and the like. In order not to confound the scoring and interpretation of scores on the writing samples obtained for this research study, all subjects would be expected to write on the same topics. In addition, all subjects would respond to more than one writing stimulus, writing on four different topics in randomized order to control for order effects, in each of two different modes of discourse (compare/contrast and chart/graph).

Pretesting of Writing Tasks

Twenty-three students in English Language Institute classes at the University of Delaware responded to a survey to obtain their reactions to 22 essay topics (10 chart/graph, 12 compare/contrast). The students represented a variety of language backgrounds and major fields of study at the university. They were given numbered examples of the 22 topics and asked to assign two ratings to each topic: (1) how difficult it would be to write an essay on the topic (1-5 range, 1 as difficult, 5 as easy), (2) the reason or reasons the topic might be difficult to write about (choices of grammar, ideas, vocabulary). They also were asked to write the number of the chart/graph topic and the number of the compare/contrast topic they would most like to write about. In addition, the research staff met with the ESL instructors to obtain their reactions to the topics and their suggestions for revisions or additional topics. The instructors supplied valuable insights regarding the different cultural perspectives of their international students and the design of writing tasks that would be the most appropriate to the objectives of the study. The international students at the University of Delaware reacted to the following topics:

Chart/Graph

1. Individual consumption of major foods in the U.S. (line graph)
2. Factors in the choice of a graduate or professional school (bar graph)
3. Planned fields of study of college seniors (pie chart)
4. Changes in automobile part production by three companies (bar graph)
5. Expenses for one family (pie chart)
6. Area and population of continents (bar graph)
7. Average height of boys and girls from birth to age 20 (line graph)
8. Changes in farming in the U.S.: 1940-1980 (bar graph)
9. Area and population of continents (two pie charts)
10. Factors in the choice of vocational field (bar graph)

Compare/Contrast

11. Travel and reading are two ways of learning about people and the world.
12. Food as a necessity vs. food as a source of beauty and pleasure

13. Potential and limitations of organizations in promoting international relations

14. Methods of decision making--careful thinking vs. quick decisions

15. Deciding between a job that pays well, but offers little enjoyment, and a job that pays less but is very satisfying

16. Advantages and disadvantages of exploration of outer space

17. Occupational preferences for working with other people vs. working by oneself

18. Advantages and disdvantages of using chemicals to control insects

19. Advantages and disadvantages of a common international language

20. Preference for spending free time in active, physical recreation vs. participation in intellectual activities

21. Advantages and disadvantages of the automobile

22. Advantages and disadvantages of very large vs. small universities/colleges

The chart/graph topics were purposely designed to present data in different forms--bar graphs, line graphs, and pie charts--in order to explore possible differential reactions to these stimuli. In fact, the ESL instructors sense that students have varied degrees of experience with data presentations; for example, students from the Middle East seem to be more comfortable with tabulated data than with graphs and charts. Topic 9, in fact, presents tabulated material juxtaposed with the pie charts for this reason. The student and instructor reactions thus guided the selection of chart/graph topics that would not create a problem in understanding the stimulus.

Eight topics, four chart/graph and four compare/contrast, were selected on the following basis: students perceived them to be of an average range of difficulty (2-4); students reported fewer reasons, particularly in regard to ideas or vocabulary, for having difficulty with the topics. Their selection of topics they would most like to write about also was considered in eliminating less promising topics. The students' preferences for the eight topics selected are summarized as follows:

Chart/Graph

1. Individual consumption of major foods in the U.S. (line graph)--overall difficulty rating of 4; reasons for difficulty--ideas (D. Food)*

4. Changes in automobile part production by three companies (bar graph)--overall difficulty rating of 3 or 5; reasons for difficulty--ideas (C. Automobile)

8. Changes in farming in the U.S.: 1940-1980 (bar graph)--overall difficulty rating of 2 to 4; reasons for difficulty--particularly ideas and vocabulary (B. Farming)

9. Area and population of continents (two pie charts)--overall difficulty rating of 3, with a few 4s and 5s; reasons for difficulty--no perceptions of reasons for difficulty (A. Continents)

Compare/Contrast

11. Travel and reading are two ways of learning about people and the world--overall difficulty rating of 4; essentially no perceptions of reasons for difficulty, though few selected grammar (3. Learning)

12. Advantages and disadvantages of exploration of outer space--overall difficulty rating of 3, with a few perceptions of difficulty with ideas and vocabulary (4. Space)

18. Advantages and disdvantages of using chemicals to control insects--overall difficulty rating of 2 and 3, with a few perceptions of difficulty with ideas and vocabulary (2. Chemicals)

20. Preference for spending free time in active, physical recreation vs. participation in intellectual activities--overall difficulty rating of 2 to 4, with a few perceptions of difficulty with ideas (1. Recreation)

Pilot Testing of the Eight Topics

Colleagues who are involved in ESL instruction offered to assist with the pilot testing of the writing sample topics. Individuals at seven different institutions of higher education throughout the United States administered the writing prompts during regularly scheduled class periods in August and September of 1983.

The samples were administered primarily to students who were preparing to enter the institutions in the fall, both at undergraduate and graduate

---

*A brief "title" for each of the topics is included in parentheses, to simplify discussion about these topics later in the report. The letters or numbers were assigned to the topics for the pilot test essay reading.

levels of education. At some schools, all eight topics were administered, whereas other schools selected particular topics they wished to use. Some students wrote on more than one topic, but most of the writing samples obtained were written by different students. The individuals who administered the writing prompts were instructed to attempt to obtain the samples under standardized testing conditions, giving students 30 minutes to write on each topic.

A total of 447 writing samples were obtained; an additional 30 that had been collected by one institution could not be used because they arrived too late to be included in the reading session. The numbers of writing samples obtained for the chart/graph topics were as follows: 46 samples for the topic labeled Continents (A); 56, Farming (B); 33, Automobile (C); and 52, Food (D). For the compare/contrast topics, the numbers of writing samples obtained were as follows: 42 samples for the topic labeled Recreation (1); 53, Chemicals (2); 100, Learning (3); and 65, Space (4).

Reading of Pilot Test Writing Samples

Prior to the reading of the pilot test writing samples, four project staff members, including the project directors and two ETS experts on scoring writing samples, read the papers to select examples to be used as range finders during the reading of the entire set of 447 samples. This sample picking required the staff members to read, and score independently, nearly the entire set of writing samples, a process that required two full days of reading. The objective was to obtain a set of writing samples that illustrated the full range of writing ability, demonstrated characteristic problems in scoring, and called attention to typical reader pitfalls (e.g., assigning a low score to a short paper). The samples selected were those for which two staff members agreed exactly on the score. A third person read and scored unusual papers. The reading of the writing samples was done holistically to obtain one score to reflect the overall impression of the quality of each paper. On the advice of Gertrude Conlan, an ETS staff member with considerable experience in scoring writing samples, the project staff decided to try out a six-point score scale. As the papers were read, they agreed that the six-point scale worked very well; a four-point scale would not have discriminated well among the set of papers obtained, and a more extended scale appeared to require more distinctions than could have been made with confidence.

Since the numbers of writing samples obtained for each of the eight topics were not large enough to justify separate training for each topic, the staff decided that the four graph/chart topics would be read together, as would the four compare/contrast topics. At the conclusion of the staff readings of the writing samples, two sets of range finders were selected to use in the training of readers for the pilot test samples. For the compare/contrast topics, 25 writing samples were selected; for the chart/graph topics, 24. These two sets of range finders were duplicated and placed in random order. For easy reference during reader training, each range finder was labeled alphabetically to correspond to this order and also labeled either alphabetically or numerically to correspond to the specific topic to which the student had responded in each writing sample. For example, the first writing sample in the set of compare/contrast range finders was labeled A, which was followed by the number 1 to designate that the topic was topic 1, Recreation. The range finders then were duplicated so that each reader would have copies. The covers had been removed from all papers previously to avoid influencing the readers with information about the writer's nationality or major field of study.

In preparation for the reading of the pilot test writing samples, it was determined that, at a rate of 35 papers per reader per hour, with two readings per paper, six readers could read the 447 papers. This estimate allowed for the possibility that ESL papers would take slightly longer to read than papers written by native speakers of English. It also included time for training, for introducing each new topic, and for discussion about each topic. The staff decided to conduct two reader training periods, one using the mixture of range finders on compare/contrast topics and one using the mixture of range finders on the chart/graph topics.

The reading was conducted on September 7, 1983. Each reader received a folder containing copies of the topics and the two sets of sample papers. The readers were instructed that the major objectives were to read the papers with the perspective of selecting the topics that "worked" best--those that showed evidence of a broad range of writing ability, that elicited the kind of writing intended, and that allowed readers to make clear distinctions in assigning scores--and to evaluate the efficacy of the six-point scale. Their objective at the conclusion of the day was to determine which two of the compare/contrast topics and which two of the chart/graph topics would be used for the large test administration. At the conclusion of the introductory remarks, which also outlined the schedule for the day, the chief reader began the training period.

The training of readers

The first part of the day was devoted to the compare/contrast topics. The chief reader asked the readers to read, score, and rank order a set of range finders, eight papers that she had selected from the samples for this topic type. Although the range finders were drawn from all four topics of each topic type, the readers scored all of the papers on one topic at a time. In reading the papers, the readers were instructed to read each paper quickly from beginning to end, to obtain an overall impression, and then to score the paper on a scale of one to six, with a score of one for a poorly written paper that minimally addressed the topic and a score of six for a top paper within the group. The readers were cautioned to avoid being inappropriately influenced by the following features of the writing samples: neatness, handwriting, occasional spelling or plurality errors, a paper with a good introduction that gradually deteriorates in quality, and a paper with a good closing statement that may or may not make up for previous weaknesses.

When the readers had finished these papers, their scores were tallied on a chalkboard. After discussion, six more papers were introduced, read, and discussed; then a few more were used to assist the readers in refining their definition of papers that should receive the midpoint scores of three and four. The training took approximately one and one half hours. The afternoon training, which observed similar procedures for the chart/graph topics, was completed in approximately one hour.

The scoring of papers

Although sufficient time had been allotted for reading the pilot test writing samples, it became clear during the morning of the reading that not all papers would be read. Papers on two compare/contrast topics were read in the morning, and the remaining two compare/contrast papers and four chart/graph papers in the afternoon. Because of time constraints and the high interreader agreement on the compare/contrast topic papers, the papers on the chart/graph topics were read by only one reader. The project staff conducted spot check readings throughout the day to ensure that the readers were scoring accurately and reliably. To resolve the very few discrepancies in scoring, one staff member read these papers a third time.

Throughout the reading, two staff members distributed the papers to readers and collected them as they were read. The aide recorded the reader number and score, covered the first score with a black sticker, and sorted the papers so that each paper would be read for the second time by a different reader. After recording the second score, the aide gave those papers requiring a third reading to the designated staff member.

After the reading of the papers for each topic concluded, the two project directors conducted discussions about the merits of each topic. The readers evaluated a topic in relation to the others of its type and suggested specific revisions. At the end of the day, the readers contributed final recommendations for the four topics that appeared to yield writing performance that met the objectives for the research and met the criteria for effective topics.

As the readers commented on the scoring, they referred to the papers used in the training, and concentrated on the following evaluation criteria, with a focus on the efficacy of the topics:

- Did the writers understand the topic?
- Did the writers address the topic directly, and did they follow instructions?
- Were the papers written in response to the topics appropriately varied in approach, or do they show the topic to be too broad or too constraining?
- Were the writers able to conclude their papers effectively?
- Did the chart/graph form of presentation present readability problems?

Conclusions and implications for the January reading

This reading session served as an opportunity for a "dry run" in preparation for the final reading of the large sample of papers in January. This experience provided concrete information about the final design of the test booklets to facilitate scoring and the mechanics of the essay reading process. The project staff decided, as a result of this experience, that the January reading would require the services of the Essay Reading Office staff at ETS to organize and run the reading procedures. Although the aide for this reading kept the papers flowing well, the need for an adequate number of aides was clearly evidenced during this reading.

With regard to reading rate, it was concluded that the original estimate of 35 papers per hour per reader was very accurate—the time pressure experienced during this reading resulted from the extended discussion that was necessary as the topics were thoroughly examined. The number of readers needed was estimated accurately, as well. If some method other than holistic scoring were to be used, the estimated reading rate would need to be adjusted.

The training time, especially in the morning, was longer than anticipated. Since different methods of scoring were planned for January, it was agreed that the training would be more complicated. Because thorough training promotes reliable reading, sufficient training time should be built into the schedule. In addition, planning would need to anticipate stretch breaks for the readers and the availability of additional sample papers to keep readers on target during the course of the reading.

The holistic scoring method worked very well. At this point, the staff anticipated the need to plan for the possibility of a two-score reading to contast mechanics/grammar and organization/coherence features without sacrificing reliability and efficiency. The readers agreed that this scoring system would be worth attempting as an addition to holistic scoring.

The readers and ETS participants agreed that the six-point scale worked very well on these papers, because sufficiently fine discriminations could be made among the papers. The six-point scale may prove useful in the future for making judgments if the essay component becomes operational in the TOEFL.

Final Selection of Topics

Two compare/contrast topics, Recreation and Space, were selected for the large-scale pretest administration at the conclusion of the discussion of the four topics of this type.

The two chart/graph topics, Farming and Continents, were selected by the readers as a result of their analyses of the four chart/graph topics.

Formatting of the Test Booklet

The physical layout of the writing stimulus requires careful consideration because the format in which the writing assignment is presented provides cues to the writer that will influence his or her response to the topic. Besides the instructions, other nonverbal cues can affect performance. The pretesting and pilot testing experiences influenced the decisions about design of the test booklet.

The cover of the booklet clearly indicates the total time allocated to writing, as well as the time limits for each topic. The thirty-minute time limit proved to be adequate for most students who wrote responses to the pilot test topics. Clearly this amount of time does not allow for much in the way of prewriting activities or revision. A longer time limit may be desirable, but we could not expect international students to spend more than two hours writing, in addition to the time taken for administrative procedures and rest breaks, particularly since each student would be writing four papers. In the event that a direct measure of writing ability becomes a section of the TOEFL or GRE General Test, score users will need to be informed that scores for the writing samples should be interpreted in the context of restricted time limits and testing conditions. A score on a writing sample administered in a testing situation cannot be assumed to represent accurately how the student would perform under optimal conditions for writing. However, since this study's research conditions should parallel the conditions of an actual testing administration, time limits that realistically suited this purpose were used.

The booklet cover also requests information for identifying the student as a research subject--name, TOEFL application number, native country, major field of study, number of years studying English, level of education, and sex. The general instructions on the cover were designed to communicate expectations regarding administration procedures; the objective of the assessment ("how well you can write"); the criteria for evaluating the writing (clear and effective expression of thoughts, emphasis on quality vs. length); and the physical presentation of the composition (more than one paragraph, writing on every line, a space for making notes). This cover was removed prior to the scoring sessions to avoid influencing readers with background information.

To eliminate the effects of the order in which the essays appeared, the four essays were assembled in eight different orders in the booklets; thus different students would be responding to different topics at the same point in time during the administration. The ordering of the topics was not entirely random, since topics of one type were not presented consecutively to minimize the possibility that the writer, having responded to a compare/contrast topic, would fall into a pattern of responding that would not be appropriate to the subsequent topic (i.e., the writer would be responding to the mode of discourse rather than to the new subject matter).

Each topic also was printed in a different color, which eventually facilitated record keeping and scoring procedures. These colors frequently were used in referring to papers written on a specific topic, and were especially useful when readers were recording their scores on the back of each test booklet.

The back cover of the test booklet was designed to allow for several different scoring contingencies. Spaces are allocated for scores assigned by two readers for each paper, for holistic and two-score methods, and for a third reading, if warranted. The size of the rectangles in which the scores would be entered corresponded to the size of the stickers that would be placed over a score assigned by the first reader of a paper. For a large-scale essay scoring operation, machine-scannable score sheets can be designed instead. We did not develop a machine-scannable score sheet, however, since the initial expenditure is considerable, though a worthwhile investment for a continuing test program.

Selection of An Indirect Measure of Writing Ability for the GRE Sample

One of the objectives of this research was to investigate the relationship of scores on indirect measures of writing ability (multiple-choice writing tests) with scores on direct measures of writing ability (writing samples). The sample of subjects in the study was to be composed of two groups: (1) international students (graduate and undergraduate) who are nonnative speakers of English and are taking the TOEFL examination; and (2) United States entry-level graduate students who are native speakers of English and are taking the GRE General Test, all as candidates for admission to institutions of higher education in the United States.

For the international candidates, Section 2 of the TOEFL, Structure and Written Expression, served as an indirect measure of writing ability. This section is composed of two parts. The first part contains items that measure the understanding of basic grammar and syntax; the items are of the "sentence correction" item type. The second part, consisting of "usage" items, tests knowledge of the grammar and usage of written English and has been demonstrated to have a consistently high correlation with writing ability (Pike, 1979; Pitcher & Ra, 1967).

For the United States candidates, however, the GRE General Test does not contain an indirect measure of writing ability. Thus, in addition to responding to the four writing sample research instruments, these candidates would have to take a separate indirect writing test. The project staff selected sections of a standardized test previously administered as a part of the Law School Admission Test (LSAT). These sections have been discontinued and disclosed, since LSAT candidates now respond to a writing sample as a direct measure of writing ability. The LSAT indirect writing test is appropriate to this sample, in that it was developed for students who have completed undergraduate education and are candidates at the graduate level. Further, its item types are parallel to the item types in the Section 2 of the TOEFL examination. Section 5 of the old LSAT was composed of sentence correction items, Section 6, of usage items.

The particular form of the LSAT indirect writing measure was selected from among the five forms that were administered during the last year it was in use. In consultation with a test development specialist at ETS who is familiar with the LSAT, we chose Form 3BLS4 on the basis of the following criteria: the mean difficulty (delta) for the items is approximately 12, an appropriate level of difficulty for GRE candidates; the items represent a good range of deltas; and the form contains only a few items that have low biserial correlations (at or below .30). In addition, the specific items in this test form have good face validity and do not contain content that is culturally biased. The Law School Admission Council granted permission to the project to use the test form for research purposes.

Since this form was originally contained in a complete LSAT examination, Sections 5 and 6 were slightly redesigned to create a nine-page test composed of a total of 60 items. The instructions that had appeared on this section of the LSAT were essentially used verbatim, with only minor modifications because the test would not be administered as a part of a larger test. The same time limits, which had been reasonable (data indicate that it was not speeded) for LSAT candidates, were retained.

Development of the Essay Reader Questionnaire

The reliability and validity of methods used for scoring writing samples is influenced strongly by the readers who apply these methods. As they evaluate samples of writing, readers are making judgments that are conditioned not only by training at the time of the essay reading but also by their personal perspectives with regard to their definitions of "good" writing and the evaluation of writing ability. One of the objectives of this research was to determine whether readers who represent different academic points of view would score the writing samples differently; thus the large-scale essay reading session was designed to involve equal numbers of readers with experience in English and ESL instruction and to have each paper read by readers from both disciplines. Information regarding the agreement of readers, such as interreader reliability coefficients, would provide some evidence of agreement among these readers. High reliability coefficients indicate that, when subjected to training in scoring methods, different readers assign essentially the same scores to the same paper. If interreader correlations are moderate or low, the reason(s) for their disagreement should be investigated.

When high interreader reliability coefficients are obtained, this may have two explanations: (1) the training sessions enabled readers, who may or may not have common views, to agree on common criteria for evaluation, and/or (2) despite the training, readers, especially those who are involved in the field of writing (English or ESL), tend to agree on criteria for evaluation. With low or moderate interreader reliability coefficients, other questions are raised with regard to the following: (1) how and if the training sessions could have been improved to obtain higher agreement; (2) what readers perceive to be their personal criteria for good writing; and (3) whether the personal criteria held by readers are significant to the evaluation of writing and should have been taken into account during the training. To gain information about readers' points of view, a reader questionnaire was designed. The staff decided that readers would be asked to repond to the instrument at the conclusion of each day of essay reading, rather than prior to the readings, to avoid heightening reader sensitivities with questionnaire prompts about evaluation criteria.

The reader questionnaires differed slightly on the two reading days because readers would be asked to react to and compare the two different scoring methods used during each session at the conclusion of the second day of reading. The questionnaire for the first day, Saturday, attempted to learn about an individual reader's criteria for evaluating writing skills with three types of items: (1) the features of writing that are valued in actual practice outside the formal reading session (e.g., in the classroom, writing center), (2) the features of writing that influenced the evaluation of writing samples during the essay reading, and (3) reactions to the scoring system used during that day of essay reading. The writing features to which readers would be asked to respond consisted of identical lists of features for the first two questions. This list of features is nearly identical to the list that was used as a part of the questionnaire in our previous survey of academic writing skills (Bridgeman & Carlson, 1983), with the addition of one feature, "mastery of the conventions of grammar." The Sunday questionnaire omitted the features to be evaluated in question 1, but again asked about the features of writing that influenced the reader's scoring during the second day of reading and for reactions to the scoring method used that day (two-score). In the final section of the Sunday questionnaire, the reader was asked to evaluate the scoring methods used on the two days. Throughout both questionnaires, the readers were given an opportunity to supply comments as well.

Therefore, the Saturday and Sunday questionnaires should provide more detailed information about points of view with regard to the evaluation of writing ability, such as the following:

- Which features of compositions are most highly valued in judging the quality of writing?
- Which features are relatively unimportant to judging the quality of writing?
- Do academics from the different disciplines, ESL and English, place different emphasis on the features of compositions as they evaluate them?
- How well do the criteria held by the readers match the explicit or implicit criteria employed in the reader training sessions?
- Can we assume that the training of readers influenced--reinforced, altered, or diminished—their personal criteria for the evaluation of writing?

- Do the criteria used for scoring writing samples during a formal essay reading have relevance to criteria used in the classroom?
- How can essay scores on a standardized test such as the TOEFL or GRE examinations be reported meaningfully (a question of appropriateness or validity)?

III. ADMINISTRATION OF EXPERIMENTAL TESTS

While the topics for the direct measures of writing ability were being developed, pretested, and pilot tested, arrangements were being made to administer the experimental instruments at TOEFL test sites and at institutions in the United States.

International Administration of Writing Samples

The project staff worked closely with the TOEFL program staff to identify international TOEFL testing centers with sufficient volumes of candidates from which to draw the research samples in countries in which Arabic, Chinese, and Spanish were the native or primary languages. Test centers in eight countries--two Chinese-speaking, three Arab-speaking, three Spanish-speaking--were selected initially; after preliminary contacts were made regarding data collection, two additional Spanish-speaking centers were added.* The TOEFL test center supervisors or agents at these sites received letters inviting them to participate in the research study during the late summer and fall of 1983. The letters contained information briefly describing the project and test administration procedures; the minimum and maximum numbers of TOEFL candidates, by levels of education and major fields, that were our sampling objectives for the site; suggested procedures for identifying or selecting candidates; a request for the supervisors' recommendations for candidate remuneration; and the form of reimbursement for their services. All the test center administrators who were contacted agreed to participate in the study.

A project objective had been to obtain all writing sample data at the time of the November TOEFL administration; however, following this administration, further testing at certain sites was scheduled for the January TOEFL administration as well. An important requirement for the data collection was that the administration of the experimental writing samples take place as close in time as possible, preferably on the same day as the TOEFL examination. This requirement was imposed so that both the TOEFL scores and the essay scores would be collected when a candidate was at a particular level of English proficiency; since English language proficiency is a developing ability, scores obtained on each measure at different times would not be comparable because they would be confounded by intervening experiences with the English language. Most test centers did administer the writing samples on the afternoon of the TOEFL examination, following a lunch break; one Arabic center administered them on the preceding day.

---

*Hereafter in the text, testing centers with a specific native or primary language will be referred to as Arabic, Chinese, or Spanish.

Another important objective of the study was to obtain data from a subsample of TOEFL subjects who had recently taken, or planned to take in the near future, the GRE General Test. Some centers were able to match candidates who had registered for the November and January TOEFL examination with a list of students in their vicinity who had taken, or had registered for, the GRE; these lists of GRE candidates were either prepared at ETS and sent to the test center administrator to do the matching, or the staff at AMIDEAST in Washington, D.C. prepared a list that matched candidates, or the administrator had the information on GRE candidates in order to do the matching (particularly if he or she also served as GRE administrator). This requirement proved to be a difficult one to meet—all international GRE candidates do not necessarily take the TOEFL. The test centers worked very hard to meet this objective, but were not able to identify and test as many TOEFL/GRE candidates as had been anticipated.

In addition to meeting the objectives of continuity of administration of the TOEFL and the collection of writing samples and for graduate-level TOEFL candidates with GRE scores, the administrators at each location were asked to meet the following criteria for the selection of research subjects: minimum/maximum numbers of candidates to be tested, the subjects taking the TOEFL at a point when they are ready to apply or are prospective candidates for admission to an institution of higher education in the United States, and subjects whose primary language is that of the country in which they would be taking the TOEFL. We also recommended that the administrators invite more candidates than were required, to allow for attrition, and that the subjects be paid on the same day that they wrote their writing samples.

The TOEFL test centers that participated in the data collection are as follows:

Arabic

Cairo, Egypt
Amman, Jordan
Kuwait, Kuwait

Chinese

Kowloon, Hong Kong
Taipei, Taiwan

Spanish

Bogota, Colombia
Santiago, Chile
Mexico City, Mexico
Lima, Peru
Caracas, Venezuela

As the most appropriate arrangements at each location were worked out, there were some variations in procedures. The following specifics varied across sites: the administrator who planned and carried out the testing (TOEFL agents, supervisor, proctors); procedures for identifying and contacting candidates; the amount of remuneration in local currency or dollars that was attractive to candidates; the scheduling of the essay administration; and the numbers and categories of subjects to be obtained. Prior to the test administrations, the following materials were mailed to each center: test booklets, supervisor's instructions, and subject consent/receipt forms.

United States Administration of Direct and Indirect Measures of Writing Ability

Data were collected at the following institutions of higher education in the United States: Rider College, New Jersey; Rutgers University, New Jersey; Southern Illinois University, Illinois; the University of California at Los Angeles, California (UCLA); and University of Southern California, California. The campus representatives experienced extraordinary difficulties with obtaining subjects—each attempted at least two administrations, but were unable to obtain as much data as planned. Candidates were offered $20 each for participating in the study. However, this stipend was not sufficiently attractive; it appears also that students in the United States are not very willing to spend their leisure time writing four papers in a testing situation.

Description of the Sample

We anticipated the possibility that the obtained score patterns would differ across language groups. The organizational style and nature of grammatical errors of Chinese-speaking students might be different from those of Spanish-speaking students, and the relationships of essay scores with TOEFL and GRE scores also might vary by language group. Indeed, Pike's (1979) findings indicate the existence of such language group differences. Therefore, we planned to study three different major language groups with large numbers of TOEFL and GRE candidates.

In any data collection, some of the data may be unusable for a variety of reasons, but very few of these cases occurred in this sample (Table III-1).

> In the Arabic language group, a total of 154 tests were administered: 44 in Egypt, 63 in Jordan, and 47 in Kuwait. However, 14 of these tests were written by students for whom Arabic was not their native language; these were omitted, resulting in a sample of 140 Arabic-speaking students, composed of 45 undergraduate- and 95 graduate-level candidates.
>
> In the Chinese language group, a total of 232 tests were administered; 69 in Hong Kong (including three graduate-level candidates), and 163 in Taiwan (including 22 undergraduate-level candidates). All 232 tests were usable, resulting in 88 undergraduate- and 144 graduate-level test booklets.

In the Spanish language group, a total of 216 tests were administered; 35 in Chile, 12 in Colombia, 42 in Mexico, 117 in Peru, and 10 in Venezuela. Venezuela and Colombia were unable to recruit as many subjects as planned. Five tests were unusable in this group, four because the subjects did not speak Spanish as their primary language. Thus for the Spanish language group, a total of 211 booklets, composed of 69 undergraduate- and 142 graduate-level writing samples, contributed to the analysis.

In the English language group, a total of 60 tests were administered; 2 at Rider College, 3 at Rutgers University, 36 at Southern Illinois University, 16 at UCLA, and 3 at the University of Southern California. Five of these booklets were not usable—two, because they were incomplete, and three, because the primary language of the writers was not English. A total of 55 graduate-level booklets for the English language group resulted. Because administrators had difficulties obtaining participation, this sample of papers may not be representative, in that students who felt that they were not capable of producing four writing samples chose not to participate.

The total number of test booklets collected was 662. Three of these booklets were incomplete, however; the remaining 659 test booklets were scored during the essay reading sessions. From this sample, 21 booklets were removed from the analysis because they did not represent the primary language of the country in which they were collected; the primary languages of the writers were an assortment of other international languages. Thus the total number of essay booklets that were used in the data analysis was 638—211 by Spanish, 232 by Chinese, 140 by Arabic, and 55 by English candidates. The total number of essay booklets written by graduate-level candidates with GRE scores was 165: 59 Spanish, 88 Chinese, 18 Arabic, and 55 English. The groupings of candidates by major fields is reported in Table III-1. The total sample is relatively representative of language groups, major fields, and graduate-level candidates with GRE scores.

Table III--1

Sample Description

Language Group	Undergraduates	Graduates				Total
		Business	Hard Science/Engineering	Social Science	Unknown	
Arabic	45	7	64	23	1	140
Chinese	89	30	65	47	1	232
Spanish	69	39	63	33	7	211
English	-				55*	55
Total	203	76	192	103	64	638

---

* Because of the small number of native English speakers, no attempt was made to classify them separately by intended major.

IV. SCORING THE WRITING SAMPLES AS DIRECT MEASURES OF WRITING ABILITY

To compare the results of scoring the writing samples by using different scoring methods, the papers were scored as follows:

- Holistic scoring of all booklets, primarily on the first day of the essay reading weekend
- Two-score scoring of all booklets, for discourse/sentence characteristics, primarily on the second day of the essay reading weekend
- Holistic scoring of a representative subsample of the papers by subject matter experts in two major fields of graduate education
- Descriptive scoring of the features of a representative subsample of the papers using the Writer's Workbench software

Other tasks needed to be accomplished before any scoring was begun, however--preparing the test booklets, planning the weekend essay reading session, selecting samples for training during the reading weekend, and refining sample selection during the training of table leaders for the reading weekend. These procedures are described in the next section of this chapter, and the application of the four scoring methods is described in subsequent sections.

Preparation for the Essay Reading Weekend

Planning for the Reading Weekend

At the point when plans needed to be made final, all test booklets had not yet arrived, nor was all testing completed. Since additional international data collections were scheduled for January, the total number of test booklets could only be estimated. It was necessary to estimate the greatest number of booklets that might be obtained in order to invite a sufficient number of readers and to organize the readings by tables and space. Thus the staff estimated that it would be possible to obtain a total of 4,000 papers, or 1,000 papers per topic. To estimate the amount of time it would take to read these papers, this figure was multiplied times four because each paper would be read twice for the holistic scoring and twice for the discourse/sentence scoring. This resulted in an estimate of 16,000 papers to be read. Using the scoring rate that appeared to be reasonable during the reading of the pilot test writing samples, that of 35 papers per reader per hour, and the actual amount of reading time that could be planned for one weekend (minus training), it was determined that 48 readers would be required.

To balance the academic perspectives of the readers, the staff decided to invite 24 ESL and 24 English readers who had experience with evaluating compositions. The number of table leaders (eight) was determined by dividing the 48 readers into tables of six; this number of readers was recommended on the basis of considerable experience with essay readings at ETS. Further, it was determined that two chief readers, eight aides, and four members of the project staff would be needed when training, space arrangements, paper flow, and the like were taken into consideration.

Sample Picking Sessions

The objective of the sample picking sessions was to select papers that represented the range of the six-point score scale, both for the holistic scoring and for the two-score scoring of the papers for each of the four topics. After a sufficient number of papers had been selected for one topic, they were arranged in an order that would be used for discussion at the table leaders' meeting, when specific papers would be selected as the benchmarks for training readers. The order in which they were arranged did not correspond to the sequence of holistic scores, but rather to a random sequence that would not suggest some predetermined score.

On the second day of sample picking, the selection of sample papers for the holistic scoring of the four topics was completed. The selection of papers for the two-score ratings for discourse-level and sentence-level (and below) skills proceeded much more slowly. As the staff and chief readers attempted to read the papers to arrive at a general impression of two scores, agreement was difficult to reach, and it was nearly impossible to not reread the paper before assigning one or both scores. Much more discussion was required before criteria for each of the two scores could be clarified. Clearly, many of the features that influence the evaluation of discourse-level skills also influence the evaluation of sentence-level skills; thus it was difficult to attempt to rate the two levels independently. Eventually, the readers were able to arrive at reasonable agreement on the sample papers for two of the topics, one chart/graph (Farming) and one compare/contrast (Space) topic. Everyone expressed concerns that the readers would experience the same difficulties, and that the criteria would be less salient, resulting in less justifiable and less reliable criteria.

Since sample papers for two topics for the two-score method had not been selected during the two days of sample picking, the project staff spent an additional day selecting the sample papers for these topics (Recreation and Continents). The staff appeared to experience somewhat less difficulty with determining these range finders. From an operational testing situation, the effort required to identify range finders for the two-score method is not efficient, unless the data demonstrate that the two scores can provide independent information about writing ability.

Given the time involved in acquiring the criteria for making the two-score distinctions, the staff decided to formulate an alternate plan for the second day of the essay reading weekend. Although it might have been possible for the readers to assign two scores to the papers written in response to the four topics, readers should not be placed under unrealistic time pressures. Rather than sacrifice reader accuracy and reliability, the staff decided that the readers would be expected to assign the two scores to papers written on only two topics, one of each type, on the second day of the weekend reading. The range finders were prepared for the contingency that these readings would proceed more quickly than anticipated, however.

All range finders, arranged in random order, were labeled sequentially with letters of the alphabet and numbers designating the specific topic (1-4) for easy reference during training and discussion. They were assembled by the Essay Reading Office and printed on colored paper to correspond with the colors of the writing stimuli as presented in the test booklets.

Chief Readers' Meeting

A table leaders' meeting was scheduled for the Friday evening prior to the essay reading weekend. On the afternoon preceding this meeting, the project staff met with the chief readers to prepare for the evening meeting. The meeting with the chief readers covered the following topics: an overview of the objectives of the research study and of the weekend reading plans, details of the mechanics of the reading sessions, and the agenda for the table leaders' meeting. Several decisions were made with regard to the orientation session for the table leaders and the conduct of the weekend readings:

- The responsibilities of the individuals involved in directing the readings
- The specific functions of the chief readers
- The specific functions of the table leaders

Table Leaders' Meeting

The project staff, chief readers, and table leaders met during the late afternoon and evening of the day preceding the reading weekend. Eight table leaders had been selected by the project staff: four individuals who have considerable experience with English composition and had served as readers or table leaders for the New Jersey Basic Skills essay readings, and four individuals who have experience with ESL composition and with essay readings in other contexts. The four ESL table leaders also had served as readers during the reading of the pilot test writing samples.

This meeting covered the following topics: an overview of the research and reading plans, the mechanics of the reading sessions, the agenda for the evening, the functions of the table leaders, and the preliminary selection of sample papers to be presented to the readers for the final selection of range finders during the reader training periods. Samples were read and selected separately for each of the four topics (e.g, all sample papers for the holistic scoring of the Space topic were read and selected first). The goal of the table leaders was to select, for each topic, eight papers that would be presented to the readers to represent the points on the score scale. These range finders would represent the entire range of score points, and should be most "typical" of the scores at each point. The range finders would be selected on the basis of having the best reader agreement, legibility, and not being too unusual. A few papers that represented atypical responses to the topic also were selected to serve as examples of papers that might present problems of which the readers should be aware.

Although we had intended also to select sample papers for the two-score scoring, we had underestimated considerably the time this meeting would take. Thus, very late in the evening, we decided to select sample papers for only two of the topics, one of each type (Space and Farming) and to complete the selection during a break in the weekend readings. The same process used in selecting the sample papers for holistic scoring was used; however, the table leaders needed to arrive at consensus on two scores for each paper. During discussion, they expressed the same concerns that the staff and chief readers had experienced during the preliminary sample picking, but did not appear to have as much difficulty in arriving at consensus. Although some substitutions of papers were made, the table leaders selected essentially the same range finders that we had selected previously.

Discussions throughout the table leaders' meeting led to the agreement that the readers should be alerted to a significant concern that we had experienced, a problem of topicality. An important scoring criterion is how well the writer addresses the topic within the constraints of the testing situation. Evaluation of the papers should take into account the total context--the subjects (native and nonnative speakers of English), the testing administration, the research objectives--as discussed in Chapter I. We agreed that some papers might seem "off-center," in comparison to other papers, if the writers in some way misinterpreted the task. If so, the readers would be instructed to focus on the quality of the writing, or, if unable to do so, to refer the paper to the table leader for scoring by the chief readers. For the chart/graph topics, in particular, the content of some papers might not be supported by the data in the chart or graph, or a writer might make generalizations going beyond the data instead of dealing directly with the data. In these cases, the readers would be instructed to place emphasis on the quality of the development of the ideas rather than on their evaluation of the correctness of those ideas.

The Essay Reading Weekend

During the weekend of January 28 and 29, the writing samples were scored, with Saturday devoted to holistic scoring, and Sunday, to the discourse/sentence level scoring.

Holistic Scoring

The chief reader for the holistic scoring training described the conduct of the readings. Readers were then given their room assignments, and each of the two chief readers conducted the training for the holistic scoring of the specific topics to be read in the two reading rooms. The readings of the topics were balanced such that the Continents (pink) chart/graph papers were read during the morning in one room, while the Space (blue) compare/contrast papers were read during the morning in the other room. During the afternoon, the Farming (yellow) chart/graph papers were read in one room, and the Recreation (green) compare/contrast papers, in the other room. Thus all readers did not read the papers on all topics, but each reader scored papers on topics of the two types. Within each reading room, ESL and English readers were balanced at each table. Numbers were assigned to the readers to facilitate distribution of the test booklets, since each paper was read twice, once by an ESL reader, and once by an English reader. As determined during the table leaders' meeting, the conventions for conducting the check readings, resolving discrepancies, and other important arrangements to ensure quality control were carried out.

At the conclusion of the holistic reading sessions, all participants filled out the Saturday reader questionnaires. Since the reading concluded later than anticipated, the readers were asked to plan to discuss their reactions to the scoring prior to the readings on the next day. Saturday evening was reserved for relaxation--a very important consideration to prevent fatigue.

Discourse/Sentence Scoring

On Sunday morning, the chief readers and table leaders met to make the final selection of range finders that had not been selected for the two-score scoring method during the table leaders' meeting. While this meeting took place, the project staff conducted a discussion with the readers to elicit their comments about the holistic scoring procedures. Most of these comments also were reflected in the questionnaires. One significant reaction, which is relevant to scoring reliability, was that the readers felt that the sample training papers for all topics were good examples of the score scale; during training, they reached consensus readily.

The discourse/sentence readings began midmorning on Sunday, with a brief introduction to the scoring procedures by the chief reader, who was responsible for this training. The chief readers then conducted the training on each topic to be scored in their respective reading rooms. During this training, the chief readers recommended that the readers assign the score for discourse-level skills first and determine the score for sentence-level skills as a second decision. Next the readers in one room scored the papers on the Farming (yellow) topic while readers in the other room scored the papers on the Space (blue) topic; the papers on the other two topics were not assigned discourse/sentence scores. The readers actually assigned the two scores as rapidly as they had assigned the single holistic scores, and did not appear to experience the difficulties we had experienced when selecting and scoring the sample training papers. At the conclusion of these readings, the readers filled out the Sunday reader questionnaires.

Cleanup Readings

To complete the scoring of all papers, it was necessary to schedule two additional full days for reading papers after the weekend reading session. One of these days was devoted to the holistic reading method, and the second day to the discourse/sentence reading method.

For the holistic scoring, the reading staff consisted of readers who had participated in the weekend readings (two English and three ESL), including the chief reader who conducted the training for holistic scoring, and ETS staff members. For the discourse/sentence scoring, with fewer papers, only one English (serving also as chief reader) and one ESL reader, plus staff, were needed. The same procedures as those used for the weekend readings were carried out to ensure standardized procedures and quality control.

The cleanup readings were required to score new test booklets (60) that had arrived after the weekend reading, to score some papers for which two holistic scores had not been assigned (72), and to resolve the scores for papers with discrepant (more than two points difference) scores. The total number of discrepancies for the holistic scores was 49 and for the discourse/sentence scores, 59. Our time estimate for reading the papers was appropriate--the readers scored approximtely 35 to 40 papers per hour; the scoring sessions, of course, included time for training on the sample papers.

Subject Matter Readings

Although we had planned to ask faculty members to read samples of papers written in response to all four topics, we decided instead to ask them to read a sample of papers on one chart/graph topic (Farming) and one compare/contrast topic (Space). When we initially contacted some faculty members, they indicated that asking them to read 200 papers was reasonable, but that reading 400 papers would be too time consuming and difficult. We also chose to obtain scores from more (four, rather than two) faculty members from each of two disciplines, and for all the samples in two sets, which would permit more valid comparisons among the several readers. Thus four faculty members in each discipline, the social sciences and the hard sciences/engineering, assigned ratings to one set of papers for each of the two topics.

The papers were selected to obtain a sample of papers on each of the two topics that were representative of the full range of holistic scores for each of the four language groups and for each major field represented in each language group. We were not able to represent all scores, languages, and major fields for either topic in cases for which the full range of scores had not been assigned, however, but the distribution of papers is representative of the total sample of papers. A total of 92 writing samples were selected for the Space topic and a total of 95 for the Farming topic.

After agreeing to assign scores to the writing samples, each faculty member received a letter of instruction, copies of the two sets of writing samples, and forms on which to enter the scores. The holistic ratings, on a scale of one through six, were expected to reflect the individual's views, as a subject matter expert in his field, in the hypothetical situation in which such ratings might be used during the process of making admission decisions about candidates. The criteria to be applied to the rating decisions were to reflect "writing competence" for academic work in the discipline of the faculty member.

Writer's Workbench Descriptive Scoring

The Writer's Workbench is a computer system consisting of several programs that offer diverse text analysis features, including proofreading, stylistic analysis, and the rules of English usage. It was developed by the Documentation Technologies Group at Bell Laboratories to assist with text editing at AT&T. The system analyzes prose passages that have been keyed in on a computer terminal. It is intended to serve as a tool that has its limitations, in that its capabilities do not encompass all the complexities of writing; however, the different programs are based on what most experts would agree are the tenets of good writing, such as avoiding wordy diction and eliminating passive voice.

At Colorado State University (CSU), faculty in the English and computer science departments obtained permission to use and adapt the Writer's Workbench programs for teaching composition, as a "research exchange" (Kiefer & Smith, 1984; Smith & Kiefer, 1982). At that time, the Workbench was not on the market; it now is available through a lease arrangement. The CSU faculty modified the programs for the needs of beginning college writers and joined the 17 separate programs to run with one command. The CSU system was used in this study to obtain numerical data on a variety of separate (analytical) features exhibited by four sets of samples of papers selected from the total sample of papers collected. One representative set of papers, selected on the basis of holistic scores, language groups, and major fields of study, was chosen for each of the four topics. For the Space and Farming topics, the same set of samples was analyzed on the Writer's Workbench that was rated by the subject matter experts--92 Space papers and 94 Farming papers. For the Continents chart/graph topic, 92 papers were chosen, and for the Recreation compare/contrast topic, 90 papers.

Joy Reid, a faculty member in the ESL department at CSU, made arrangements for Roberta Scott, a composition instructor who also operates a writing service, to key in the papers on the terminal. She entered each paper verbatim and subsequently obtained Workbench analyses for the four sets of papers. The output yields an astounding amount of numerical data--with Joy's informed advice, we selected the data that would be the most dependable and meaningful in the data analyses. In instances in which the data were overlapping, such as the number of pronouns and the percentage of pronouns, we selected the percentage figures. The Style program produces a large quantity of numerical scores for the various quantifiable features of prose, whereas the Prose program supplies interpretive comments regarding many of these same features. The Prose comments compare the paper's style values against a set of standards and describe the differences to the reader. Since some overlap occurs between these programs, we eliminated that redundant data as well. Thus the Writer's Workbench system provided a considerable number of objectively derived "scores" for the various quantifiable features exhibited by each of the papers in the four representative subsamples.

Descriptions of the features analyzed by the Writer's Workbench are presented in the paper by Smith and Kiefer (1982). The specific features that became relevant in the data analyses are defined more fully in the section of Chapter V that reports the Writer's Workbench data.

Scoring of Other Instruments

LSAT Indirect Measure

With permission from the Law School Admission Test (LSAT) program, a retired form of the LSAT measure of indirect writing ability was administered to the 55 students who are citizens of the United States with English as their primary language. The test consists of a total of 60 items, 35 of the usage type and 25 of the sentence correction type. The tests were hand scored, resulting in number-right scores for the total test and for each of the two sections.

Reader Questionnaires

Most of the readers, table leaders, and chief readers completed the reader questionnaires at the conclusion of each of the readings on the Saturday and Sunday of the weekend reading session. Since the table leaders and chief readers also were involved in reading papers, their responses were combined with the reader responses. On Saturday, a total of 50 participants, 24 ESL and 26 English, completed the questionnaires. On Sunday, a total of 51 participants, 24 ESL and 27 English, completed them.

The responses to open-ended questions were recorded verbatim and are available on request. These comments were sufficiently informative, interesting, and varied that we did not attempt to categorize them. The responses that required choices among a list of alternatives were entered on a data tape for further analyses.

The responses to the reader questionnaires obtained at the conclusion of the Saturday readings consist of reader reactions to criteria used to evaluate written assignments, both in their everyday experience in evaluating writing samples (e.g., in instruction) and as they evaluated the papers during the holistic scoring. Additional questions asked for their reactions to the holistic scoring of this sample of papers. The Sunday questionnaire responses again asked the readers to respond to the same set of criteria to report how they evaluated the papers during the discourse/sentence scoring. In addition, they were asked to react to the discourse/sentence method of scoring and to compare the discourse/sentence scoring method with the holistic scoring method. These responses are subjective, of course, therefore they do not provide accurate documentation of the actual processes used by the readers as they evaluate writing samples. The data indicate, however, the readers' perceptions of the various approaches to the assessment of writing ability. These results are reported in Chapter V.

V. RESULTS

The various test scores that were obtained were viewed in several ways--descriptive score distributions, estimates of reliability, exploratory and confirmatory factor analyses, and correlational and regression analyses. The first two sections of this chapter describe the test score data for the different candidate populations in the sample and subsamples. The next section reports the estimates of reliability for the scores assigned to the writing samples. The following section reports the results of the exploratory and confirmatory factor analyses and the relationships of the test scores to the factors and other test scores. Finally, the data obtained from the Writer's Workbench analytical scoring are described. The sizes of the samples for the total sample and different subsamples are not equivalent to the sizes of the samples obtained in the data collection. They are somewhat reduced because the data that were subjected to factor analyses consist of candidate scores and demographic variables that were complete, so that no calculations were made for individuals with missing data (e.g., age not reported).

Descriptions of Scores on the Conventional Tests

TOEFL Scores

The total sample of foreign students with complete data and with TOEFL scores was 542, consisting of 138 Arabic language, 230 Chinese language, and 174 Spanish language candidates (Table 1). The mean TOEFL score for the total sample was 519.97 with a standard deviation of 64.08. The means of the section scores for the total sample on the TOEFL were relatively equivalent, rounded to a mean of 52 for each section.

The means of the TOEFL scores for the Spanish language group are only slightly higher than the means for the Chinese language group, whereas the means for the Arabic language group are the lowest. These means, when compared with the normative data for the same language groups (Table 1) as reported in the TOEFL Test and Score Manual (1983), suggest that the candidates in these samples are above average. This result was anticipated, because the sample consisted of volunteer students who claimed they anticipated coming to the United States to study within the next year and who apparently felt competent to write in English.

GRE General Test Scores

The sample of international and United States candidates who completed the four writing samples, and for whom we could obtain scores on the GRE General Test, was 172, consisting of 124 international students and 48 United States students (Table 3). The sample of international candidates who took the TOEFL was 124. The number of students in the United States sample for whom GRE General Test scores could be obtained, and who also completed the LSAT writing test, was 43.

The mean GRE verbal score (385) for the total sample (Table 2) is considerably lower than the mean GRE verbal score (471) for all examinees who took the GRE General Test between 1981 and 1983 (as reported in GRE Guide to the Use of the Graduate Record Examinations Program, 1983-84), with a larger standard deviation (145) than reported for all candidates (130). The mean GRE quantitative score (635), however, is substantially higher than the mean GRE quantitative score (537) for all candidates, with a lower standard deviation (114) than for all 1981-1983 candidates (137). These GRE quantitative scores probably reflect the level and kind of preparation of the candidates in the sample for this study, since a large number of the students indicated plans to major in the hard sciences and engineering in graduate school. In fact, the GRE quantitative average test scores of 1981-1983 examinees intending to major in the biological sciences (bioscience subtotal mean of 580) and the physical sciences (subtotal mean of 628) are higher in general than the average test scores of intended humanities (means ranging from 458 to 521) and social sciences (means ranging from 434 to 603) majors. The mean GRE analytical score (488) for this sample is slightly lower than for all examinees (501), with a slightly lower standard deviation (120) than for all 1981-1983 examinees (127). The mean for GRE analytical more closely approximates the average GRE analytical test scores (491-528) of examinees who intended to major in the social sciences. Thus we observe a somewhat different pattern of scores for the GRE General Test for this sample than would be expected if the sample had consisted of predominantly native speakers of English. Where the TOEFL score means tended to be slightly higher than for the average TOEFL candidate population, the GRE scores do not reflect this pattern because the GRE norming sample is composed mostly of native speakers of English, whereas the TOEFL is normed on an ESL group.

Table 3 compares scores on the sections of the GRE General Test obtained by the foreign and United States candidates. The mean (551.46) of the GRE verbal scores for the students for whom English is their primary language (United States) is substantially higher than the mean (320.00) for the foreign candidates for whom English is not their primary language. The standard deviation (91.61) of the scores for the foreign group is considerably smaller than the standard deviation (121.29) for the United States group as well. This result is not surprising, of course, since English language proficiency is evaluated in the GRE verbal sections. The mean (567.71) of the GRE quantitative scores for the United States candidates is considerably lower than the mean (660.81) for the foreign candidates, probably reflecting again the large number of foreign students who plan to major in, and have prepared for, the hard science and engineering fields. Finally, the mean (583.33) of the GRE analytical scores for the United States candidates is considerably higher than the mean (450.48) for the foreign candidates; however, the difference between the two groups is not as striking as the difference between means on the GRE verbal. Since the GRE analytical is considered to be confounded with verbal ability (in English), this result suggests that the GRE analytical also assesses some form of analytical reasoning ability that is not as entirely dependent on English language proficiency as evidenced in the GRE verbal sections.

LSAT Writing Test Scores

The means of the scores obtained by the United States candidates on the 60-item indirect measure of writing ability also are shown in Table 3. For both sections of the test, the candidates appeared to perform equally well on the usage items and the sentence correction items; the mean score for each section reflects approximately 60 percent correct answers. The mean score on the usage section for this sample of students was 21.05, with a standard deviation of 6.62. This mean is only slightly higher than the mean score of 20.97 obtained on this form by a population of 2300 LSAT candidates in 1979, with a standard deviation of 6.20 (internal document). The mean score on the sentence correction section for the LSAT candidates was 14.46, with a standard deviation of 4.14; performance on this section by our U.S. sample, who obtained a mean score of 14.72, with a standard deviation of 4.32, was approximately equivalent. Thus this sample of U.S. candidates performed on an indirect measure of writing ability at levels represented by a somewhat selective group of graduate-level candidates for law school.

Writing Sample Scores

The means and standard deviations of the writing sample scores are presented in Tables 3 and 4.

Means and Standard Deviations--Foreign Sample

To facilitate cross-task comparisons, only subjects with complete data on both the writing samples and the TOEFL were included in these analyses. The writing samples were assigned ratings on a one through six scale. The means of the scores reported in Table 4 were averaged over two readers. For every writing sample score, the means are lowest for the Arabic sample, in the middle for the Chinese sample, and highest for the Spanish sample. When the holistic score means are compared with the discourse/sentence score means, the two scoring methods essentially yield the same mean levels of performance. In addition, except for level difference between language groups, the mean writing sample scores for the different topics are approximately equivalent. This result suggests that (1) the different topics did not elicit qualitatively different writing performance, and/or (2) the readers maintained a comparable scale for evaluating the writing samples, despite possible performance fluctuations from topic to topic.

Means and Standard Deviations--English-Speaking U.S. Sample

The data summarized in Table 3 also include only subjects with complete data. This table compares the score differences between the international and United States candidates. The mean (20.53) of the holistic writing sample scores for the English-speaking group is considerably higher than the mean (12.56) for the foreign group, but the scores have approximately the same standard deviation. Thus the average score on papers written on one topic for the United States candidates is five; for the international candidates, three (on the scale of one through six).

The means of the discourse/sentence scores for the United States group also are higher than the means for the foreign group and with approximately equivalent standard deviations. Thus the average score on papers written on one topic for the United States candidates is five; for the foreign candidates, three. The two scoring methods clearly did not yield different evaluations of the average level of quality of the papers for these two groups.

Estimates of Score Reliability for Writing Samples

Reliability of Holistic Scores

Reliability coefficients reflect the extent to which a test provides consistent results. The reliability coefficient is a generic term. Different reliability coefficients can be based on various types of evidence, with each type of evidence having a different meaning. For the current study, evidence for the reliability of the scores on the direct measures of writing performance involved the several sources of error that may reduce the reliability of scores assigned to these measures--consistency of writing sample scores across readers (raters), across topics within topic types, and across topic types.

Interrater reliability

Each paper was read initially by two readers. If the ratings assigned by the two readers were more than two points apart, the paper was read by a third reader, and the most discrepant rating was dropped. As indicated in Table V-1, there were relatively few cases where the readers were more than two points apart in their holistic judgments. Out of the 2552 pairs of judgments (638 students X 4 essays each), discrepancies greater than two points were found in only 56 (2.2%) of the cases. Correlations between the ratings assigned by the original two readers (i.e., before eliminating any discrepant scores) are presented in Table V-2.

Table V-1

Papers with Discrepant Holistic Scores

Topics	Percent of papers with disagreement of more than 2 points between reader 1 and reader 2
Compare/Contrast--Space	2.6%
Compare/Contrast--Recreation	2.4%
Chart/Graph--Continents	3.3%
Chart/Graph--Farming	2.6%

Table V-2

Interrater Correlations of
Holistic Scores

Topics	r	Spearman-Brown corrected r
Compare/Contrast--Space	.74	.85
Compare/Contrast--Recreation	.71	.83
Chart/Graph--Continents	.66	.80
Chart/Graph--Farming	.73	.84

The interrater reliabilities were consistently high for all topics and appear to represent about the best that can be expected with complex judgments of this type (Breland & Jones, 1982). The uncorrected correlation coefficient is an estimate of the reliability if only the scores from one judge are to be used operationally; if two judges are to be used, the Spearman-Brown correction provides an estimate of the reliability of the scores based on summing the judgments of two raters. Although the precise numerical impact of using a third reader to adjudicate score discrepancies of more than two points cannot be directly estimated, the values in Table V-2 may be taken as a lower bound for the reliability of the adjudicated scores. In all subsequent analyses in this section, the adjudicated scores were used. However, note that because of the small number of scores that were changed, it would make very little difference whether adjudicated or unadjudicated scores were used.

Reliability across topics

In addition to disagreements between raters, another source of inconsistency in writing sample scores may be differential student performance on different topics. Some students may find some topics easier than others, or certain topics may demand different kinds of discourse skills that also elicit differential performance. The intercorrelations among the holistic scores on the four topics are presented in Table V-3.

Table V-3

Across-Topic Correlations
Among Holistic Scores

		1	2	3	4
1	Compare/Contrast--Space		.71	.69	.73
2	Compare/Contrast--Recreation	.85		.66	.71
3	Chart/Graph--Continents	.84	.81		.68
4	Chart/Graph--Farming	.86	.85	.83	

Note: Correlations under the diagonal are corrected for interrater unreliability.

Correlations under the diagonal are estimates of what the correlation among topics would be if readers were perfectly reliable. Note that both the corrected and uncorrected coefficients indicate that correlations are no higher within topic types than across topic types. Thus, for example, compare/contrast topic 1 is not more highly correlated with the other compare/contrast topic than it is with the chart/graph topics. This suggests that, at least for these topics, there are not systematic differences in the way each topic type ranks students. The correlation of .83 between total score for one topic type (formed by adding the two scores for the topic type) with the total score for the other topic type is consistent with this suggestion. When corrected for unreliability, the correlation between the totals for the two topic types is approximately 1.0.*

Reliability within language groups

The above reliability analyses were repeated in each of the four language groups. Because of greater score homogeneity within groups, the correlations were slightly lower, but the patterns were remarkably stable. All the generalizations made about score relationships in the total group apply to the subgroup analyses. For example, the correlation between the compare/contrast total and the chart/graph total (.83 in the total sample) was .72, .75, .84, and .69 in the Spanish, Chinese, Arabic, and English samples, respectively. The above correlations estimate the reliability of half of the test. They may be corrected by the Spearman-Brown formula to estimate the reliability of the entire writing sample. The estimated reliability for all four writing samples is .91 for the combined language groups; the estimated reliabilities are .84, .76, .91, and .82 in the Spanish, Chinese, Arabic, and English samples, respectively. The reliability in the English sample is remarkably high, given the ceiling level performance of many students in this group.

---

*High estimates of reader reliabilities for holistic and discourse/ sentence scores assigned by different readers to the same and different topics indicate that readers are able to reach considerable agreement on the relative quality of a set of papers they are judging. However, this evidence does not indicate whether different readers are evaluating the same features of writing or whether they are attending to different features when making decisions to assign a specific score to writing samples that address different topics (content) and require different approaches to the task (e.g., compare/contrast vs. chart/graph). During the pretest readings, pilot test samples elicited by the different topics elicited apparent differences. Although we have no means by which to establish that the readers adjusted their standards with respect to the specific features, depending on the specific topic and its task demands, the possibility cannot be rejected. The responses of readers to the questionnaires, reported in the next section, and the Writer's Workbench analyses, summarized at the end of this chapter, offer some insights about the features of writing samples to which the readers may have been attending.

Reliability of Discourse- and Sentence-Level (D/S) Scores

One compare/contrast topic (Space) and one chart/graph topic (Farming) were scored separately for discourse-level characteristics and sentence-level characteristics.

Interrater reliability

Table V-4 presents the percent of papers on which the two readers disagreed by more than 2 points on either the discourse-level or sentence-level scores.

Table V-4

Papers with Discrepant D/S Scores

Scores	Percent of Papers with Disagreement of More Than 2 Points
Space--Discourse level	2.4%
Farming--Discourse level	2.8%
Space--Sentence level	4.1%
Farming--Sentence level	1.9%

The corresponding correlations between readers are presented in Table V-5.

Table V-5

Interrater Correlations of D/S Scores

Scores	r	Spearman-Brown Corrected r
Space--Discourse level	.66	.80
Farming--Discourse level	.72	.84
Space--Sentence level	.71	.83
Farming--Sentence level	.72	.84

The reliabilities for the sentence-level and discourse-level scores are essentially identical and are also comparable to the interrater reliabilities of the holistic scores.

Reliability across score types and across topics

Table V-6 permits comparison of the relationship between discourse-level and sentence-level scores, both within topics and across topics.

Table V-6

Correlations Between Discourse and Sentence Level Scores Within and Across Topics

		1	2	3	4
1	Space--Discourse level		.66	.87	.72
2	Farming--Discourse level	.81		.63	.86
3	Space--Sentence level	1.06	.75		.73
4	Farming--Sentence level	.88	1.02	.87	

Note: Correlations under the diagonal are corrected for interrater unreliability.

The highest correlations were across score types within topic. Thus, for example, the discourse-level score from the Space topic correlates more highly with the sentence-level score from the same topic than it does with the discourse-level score on the Farming topic. This pattern may be partially explained by the scoring strategy in which the same reader assigned both a discourse-level and a sentence-level score at the same time (which also may explain the correlations greater than 1 in the corrected correlations under the diagonal). However, it legitimately suggests that an operational program would gain nothing from a two-score system, at least if both scores are assigned by the same rater.

Summary scores were formed by adding the two discourse-level scores to form a discourse total and the two sentence-level scores to form a sentence total. The correlation of the discourse total with the sentence total was .90, further reinforcing the view that the two scores can be treated essentially interchangeably. Furthermore, the discourse total correlated .87 with a holistic total formed by adding the holistic scores on the same two essays (Space and Farming), and the sentence total correlated .88 with the holistic score. Thus, the discourse-level score, the sentence-level score, and the holistic score all appear to be assessing the same underlying writing skill.

Reliability within language groups

As with the holistic scores, the pattern of correlations within subgroups for the discourse/sentence scores paralleled the findings in the group as a whole. Correlations of the two discourse-level scores, the two sentence-level scores, and the discourse-level total with the sentence-level total for each language group are presented in table V-7.

Table V-7

Correlation of D/S Scores
Across Language Groups

Language Group	D-Level Space vs Farming	S-Level Space vs Farming	D-Total vs S-Total	N
Arabic	.60	.63	.88	138
Chinese	.54	.61	.87	230
Spanish	.54	.60	.85	174
English	.23	.18	.66	43
Total	.66	.73	.90	585

Except for the low reliability in the native English sample (where many scores were at ceiling levels), reliability of the scores was remarkably consistent across groups.

Reliability across ESL and English readers

For the interrater reliability estimates presented in Tables V-2 and V-5, score 1 was the first score assigned to the writing sample and score 2 was the second score assigned. Score 1 could be either assigned by an ESL reader or a regular English teacher reader with score 2 then being from a rater in the other group. For the analyses in this section, interrater reliabilities were recalculated so that the first score in each pair was the score assigned by the ESL reader and the second score was assigned by the English teacher reader. If ESL readers assigned scores that were systematically higher or lower than the English teacher readers, then the recalculated interrater reliabilities could be higher than the originally calculated reliabilities. However, this was not the case. As is evident in Tables V-8 and V-9, the mean scores assigned by the two types of readers were nearly identical, and the interrater correlations were very similar to those reported in Tables V-2 and V-5. Thus, the careful training procedures employed in this study were sufficient to overcome any differences in rating strategies between the two types of readers that might otherwise have occurred.

Table V-8

Means, Standard Deviations, and Correlations for Holistic Ratings by ESL and English Teacher Raters
(N=638)

Topics	Reader	M	SD	r
Compare/Contrast--Space	ESL	3.3	1.4	.67
	English	3.2	1.4	
Compare/Contrast--Recreation	ESL	3.5	1.4	.70
	English	3.4	1.3	
Chart/Graph--Continents	ESL	3.4	1.3	.67
	English	3.1	1.3	
Chart/Graph--Farming	ESL	3.3	1.4	.72
	English	3.4	1.4	

Table V-9

Means, Standard Deviations, and Correlations for Ratings by ESL and English Teacher Raters
(N=238)

Topics	Reader	M	SD	r
Space--Discourse level	ESL	3.5	1.5	.65
	English	3.5	1.3	
Farming--Discourse level	ESL	3.4	1.4	.70
	English	3.5	1.3	
Space--Sentence level	ESL	3.2	1.5	.71
	English	3.0	1.5	
Farming--Sentence level	ESL	3.3	1.5	.72
	English	3.1	1.4	

Reader Responses to Weekend Reading Questionnaires

To obtain information about the points of view held by readers with regard to the evaluation of writing skills and their exposure to different methods of scoring papers on the same topics, the readers were asked to respond to two questionnaires during the essay reading weekend. The first questionnaire (Saturday) was completed at the conclusion of the holistic scoring session. The second questionnaire (Sunday) was completed at the conclusion of the discourse/sentence scoring session.

The first section of each questionnaire consisted of a checklist of identical features relevant to the evaluation of written assignments. On the Saturday questionnaire, readers were asked to rate the degree of importance they attribute to the 13 features of written assignments in their actual practice outside of formal reading sessions (e.g., in the classroom or writing center). On the second page of the Saturday questionnaire, they also were asked to rate the degree of importance they attributed to the same features during the holistic readings. On the Sunday questionnaire, the same checklist was repeated, on which readers rated the features with regard to degree of importance attributed to the features during the discourse/sentence readings. The ratings reported by all readers who completed the questionnaires, including chief readers and table leaders, appear in Table 5. Some of the most salient responses indicated the following:

- The readers assigned high importance ratings (5) to some features they perceived they attended to, both prior to the readings and during the discourse/sentence readings: mastery of the conventions of grammar, quality of sentence structure, quality of paragraph organization, and addressing the topic. The means for these features reflect a similar pattern, although with somewhat more importance given to the features either prior to the readings or during the discourse/sentence readings. These responses suggest that the readers felt that certain features they regard as significant in practice are features to which they felt they had attributed significance during the discourse/sentence readings. In fact, their subjective reactions during discussions and conversations suggested that they perceived the discourse/sentence scoring to be more "realistic." The readers may have perceived that they were evaluating the features of the papers somewhat differently during the holistic scoring and the discourse/sentence scoring; however, the means and standard deviations of the scores for the papers do not support that the different scoring methods resulted in different levels of scores for the papers.

- The readers rated the feature, quality of overall paper organization, to be of greater importance during the discourse/sentence scoring than in other instances. This perception may have resulted from the division of ratings in the two-score method, in which discourse-level characteristics were evaluated separately from sentence-level characteristics. Overall paper organization was likely to be one of the discourse-level characteristics on which readers focused--although holistic scoring also places considerable emphasis on this feature.

- Finally, some features received higher importance ratings for the evaluation of papers prior to the reading sessions: quality of content, development of ideas, adopting a tone...appropriate to the audience, and appropriately meeting assignment requirements. These features, of course, justifiably are of more importance to classroom assignments. The ratings for these particular features probably also reflect the appreciation, which emerged during the training discussions, that the readers should evaluate the quality of the papers within the context in which they were written (e.g., time limits, possible lack of familiarity with the content of the topic, cross-cultural differences in presenting ideas and communicating tone).

Tables 6, 7, and 8 present the reader responses to the same sections of the questionnaires, criteria used to evaluate written assignments, with a breakdown comparing all readers, ESL readers, and English readers, for their ratings based on perceptions of the features prior to the reading, during the holistic scoring, and during the discourse/sentence scoring, respectively. When the perceptions of the importance of these features to ESL and English readers are compared, essentially no differences appear, as reported in any of the three tables.

Tables 9, 10, and 11 summarize the readers' responses to the questions on the two final sections of the questionnaires that focused on the scoring methods. Table 9 compares responses to the same questions, answered both after the holistic scoring and after the discourse/sentence scoring, supplied by all readers. Questions 6 and 7 appeared only on the Sunday questionnaire, since they asked readers to compare the two scoring methods. The responses indicate the following:

- Many readers (70 percent) felt that holistic scoring can be used appropriately in the classroom, but only 57 percent responded positively to the use of discourse/sentence scoring in the classroom.

- A considerable number of readers (80 percent) felt that the scores they were asked to assign during both scoring sessions were appropriate for the particular sample of papers.

- A large percentage of readers (60 percent) felt that it was possible to make clear distinctions between the papers at adjacent score intervals during the holistic scoring; however, fewer readers (45 percent) were comfortable with the discourse/sentence scoring in this regard. Many readers also informally reported the latter reaction to us.

- Only half of the readers felt that it might be possible to assign descriptions to each of the score intervals used during both the holistic and discourse/sentence scoring. The comments of many readers, both informally, and as reported in their comments on the questionnaires, indicated that they would feel uncomfortable attempting to assign descriptions to the score levels because individual papers at one score level can differ considerably but deserve an equivalent rating. The reader comments indicated, however, that sample papers at each score level could be useful and meaningful, both to other readers and to those who would interpret writing sample scores.

- Questions 6 and 7, asked on the Sunday questionnaire, reflect the readers' generally positive attitude toward the scoring methods that had been applied to the papers.

Tables 10 and 11 provide the breakdowns to the same questions, comparing the responses of all readers, ESL readers, and English readers, on the Saturday and Sunday questionnaires, respectively. The reactions of the ESL and English readers do not appear to differ and reflect the same pattern of responses as summarized above.

Correlations of Holistic Scores with Ratings of Subject Matter Experts

As noted above, the ratings of English teachers and ESL teachers agreed very well. But would judgment of professors in the substantive areas of social sciences and engineering agree with the judgments of the ESL and regular English teachers, especially if no special training for the professors was conducted? Each of four social science professors and each of four engineering professors rated (on a 1-6 scale) 90 writing samples on the Space topic. Judgments over the four professors were averaged to form a mean social science judgment; similarly the ratings of the four engineering professors formed a mean engineering judgment. The mean social science judgment correlated .86 with the holistic score that had been assigned during the regular scoring session, and the mean engineering judgment correlated .92 with the holistic score. The social science judgment and the engineering judgment correlated .92. A similar pattern was observed for the sample of 93 essays on the Farming topic that were rated by the subject matter professors. The mean social science judgment correlated .83 with the holistic score, and the engineering judgment correlated .82. The intercorrelation of the engineering and social science judgments was .92. Whatever differences in the perception of good writing may exist among regular English teachers, ESL teachers, social science teachers, and engineering teachers, these differences do not interfere with the ability of these diverse groups to rank students' writing samples in the same order.

The judges also were asked, after rating each set of papers on one topic, to indicate the rating that reflects the minimal level of writing competence acceptable for beginning students in their departments. For the Space topic, four judges indicated that a rating of 4 would be acceptable, and two judges indicated acceptability ratings of 3 and 5. For the Farming topic, most judges (six) found a rating of 4 to be acceptable, with one judge indicating a 3, and the other judge, a 2.

Exploratory and Confirmatory Factor Analyses

A series of principal axes factor analyses with varimax rotations were conducted to generate hypotheses about the factor structure of the data. The data that were factor analyzed initially consisted of the correlation matrix of the 11 variables that represented complete data for the majority of the subjects in the sample: scores on the three sections of the TOEFL, holistic scores assigned to papers on each of the four topics, and discourse/sentence scores assigned to papers on each of two topics. These analyses were conducted for the total sample of subjects (560) and for each of the three non-English-language groups, Arabic (139), Chinese (230), and Spanish (191). Several factor analyses (principle components) were conducted using the 11 variables. However, because high correlations between the holistic scores and the discourse/sentence scores indicated that the discourse/sentence scores did not represent independent information, they were omitted from the analysis. Thus, the final factor analysis consisted of the four holistic scores and the three TOEFL scores. The different analyses indicated that the data were not likely to yield more than three factors. The factor analyses of the seven variables suggested that two factors appeared to achieve a more satisfactory fit to the data. The two-factor varimax solution for the total sample accounted for 77 percent of the total variance. Subsequently, a promax factor analysis with oblique rotations was conducted, using the same data; this analysis suggested that the two factors were substantially correlated, but the promax factors did not achieve a better fit to the data than the varimax factor analysis.

The two-factor varimax solution resulted in what appear to be "method" factors (Table 12). One factor consists of the scores on the three sections of the TOEFL and the other factor of holistic scores on the four topics. The factors obtained for each non-English language group are presented in Tables 13 to 15. For this sample of data, the method of assessment appears to influence performance more strongly than the four different writing sample topics or types (compare/contrast and chart/-graph) or than the three modes of English proficiency measured by the TOEFL (listening comprehension, writing ability, reading comprehension). One interpretation suggests that performance on measures of English language proficiency becomes more differentiated when English proficiency measures require a candidate to respond by applying different cognitive processes--recognition vs. production.

Because the scores on the variables (TOEFL and holistic writing sample scores) were highly correlated (Table 16), the question still remained--whether a two-factor solution achieved a better fit to the data than a one-factor solution. A maximum likelihood factor analysis (LISREL) was conducted to determine whether the two factors would reproduce the original variance/covariance matrix. This method permits (in fact, requires) specification of a factor model of the domain to be analyzed and provides a significance test to indicate how well the model fits the data. These features of the analysis provide a rational and statistical basis for choosing the most appropriate solution from among reasonable alternatives.

The two-factor model was specified for the seven variables in the principal axes analysis. The model is revised, as necessary, on the basis of residual correlations among variables to see if a more satisfactory fit to the data can be obtained.

We limited attention to factor models having a "simple structure," allowing each score to contribute to the definition of only one factor. The first analysis (LISREL) held that the pattern of loading was invariant. This analysis showed that the goodness of fit to the two-factor solution is high (mean index of .98 over the three language groups), with a low root mean squared residual (mean of .24) that indicates most of the observed covariances in each population are explained by the two-factor model. When summed across the three language group populations, the Chi-square (42.50 with 39 df) did not reject the hypothesized two-factor solution. Next, a one-factor model was tried, but this model did not fit the data (Chi-square = 215.58 with 42 df). Although the one-factor model fit the data for the Spanish group reasonably well, it did not fit for either the Arabic or Chinese groups.

The second analysis (LISREL) assumed not only the same pattern of loadings but also that comparable loadings are equal. This solution was rejected. Again, the solution fit the Spanish group the best, but did not fit the other groups well. Taken together, the two LISREL analyses demonstrated that, for the two-factor solution, the patterns are the same for the three language groups, but the individual loadings on each factor may differ for each language group.

Relationships of Writing Sample and TOEFL Mean Scores

The model obtained by the factor analyses can be interpreted further by studying the correlational relationships between the variables contributing to the two factors and other test score variables investigated in this study.

Mean writing sample scores and TOEFL scores for the foreign samples are presented in Table 17. To facilitate cross-task comparisions, only subjects with complete data on both the writing samples and the TOEFL were included in these analyses. Writing sample scores are reported on a 1-6 scale (averaged over two raters); TOEFL scores are the standardized scores normally reported. The TOEFL total scores (reported in the last column of Table 17) may be compared with normative data for the same language groups as reported in the TOEFL manual (1983). The manual reports TOEFL scores of 463, 503, and 504* for Arabic, Chinese, and Spanish speaking groups, respectively. Thus, in each language group, the sample for the current study is above average, as discussed in a previous section.

---

*Means for the three countries contributing the majority of the Spanish speaking subjects for the study were higher. Means for Chile, Mexico, and Peru were 520, 514, and 513, respectively.

The pattern of means across the three language groups is highly consistent; for every writing sample score and every TOEFL score the Arabic sample is the lowest, the Chinese sample is in the middle, and the Spanish sample is highest. This lack of major interaction between type of score (writing sample or multiple-choice) and language group is consistent with the notion that both types of scores may be assessing, to a large extent, the same underlying language proficiency dimension. Nevertheless, there is some evidence that the between-groups differences are smaller for the essays than for the TOEFL.

In order to put both measures on a comparable scale, the writing sample holistic total scores and the TOEFL total scores were each separately standardized (z-transformed) using the mean and standard deviation of the total group for each measure. The resultant z-scores for each group are presented in Table V-9.

Table V-9

Standardized Writing Sample Essay and TOEFL Scores

	Arabic	Chinese	Spanish
Holistic Writing Sample	-.298	-.111	.380
TOEFL Total	-.636	-.067	.416

The relative positions of the Chinese and Spanish groups are essentially the same on both measures, but the Arabic sample is relatively lower on the TOEFL than on the essays. Readers who are more comfortable with percentiles should note that the Arabic group is at the 38th percentile on the writing samples but at only the 27th percentile on the TOEFL. In general, if a highly reliable measure shows some differences among groups, a less reliable measure would be expected to show less difference. However, the reliability of the writing sample total score is sufficiently high that reliability alone is probably not sufficient to explain the observed differences. Assuming that more than a statistical artifact is involved, either the TOEFL may differentially assess the language proficiency of Arabic speakers or the writing samples may be biased in favor of Arabic speakers. Further research relating both measures to external criteria is needed.

Relationship of Demographic Variables to Writing Sample and TOEFL Scores

Correlations were computed between the demographic variables and the writing sample scores and TOEFL scores in order to identify which demographic variables are significantly related to the criterion scores. The demographic variables considered were age, sex (M= 0, F= 1), undergraduate vs. graduate applicant (undergraduate= 1, graduate= 0), business vs. other graduate majors (business= 1, other= 0), hard science/engineering vs. other graduate majors, social science vs. other majors, and self-reported number of years spent studying English. The statistically significant correlations are summarized in Table 18 for the full sample of international candidates, and in Tables 19, 20 and 21 for the Arabic, Chinese, and Spanish samples, respectively. For each language group there were 91 (7 demographic variables X 13 criterion measures) correlations; thus some of the "significant" correlations may in fact be chance occurrences. Even if truly statistically significant, correlations below .25 indicate that so little of the criterion variance is explained that they have almost no practical significance.

Across all three samples, number of years of studying English is the one variable that is consistently related to all the criterion scores. Note in particular that, in each sample, the correlation with the holistic total is very similar to the correlation with the TOEFL total, indicating that years of study of English has approximately an equal impact on both methods of assessing English competence.

The correlations in the Chinese sample must be interpreted cautiously because of the split in that sample between Taiwan and Hong Kong. Most of the undergraduates came from Hong Kong while most graduates came from Taiwan. Thus, the higher scores for undergraduates (positive correlation with undergraduate status variables as well as the negative correlations with age) may be an artifact related to generally higher English competence in the British colony than in Taiwan. However, note that the higher scores for undergraduate/Hong Kong students were found consistently on the writing sample scores but not on any of the TOEFL scores. This may reflect a relatively greater emphasis on written communication skills in Hong Kong or a greater emphasis on TOEFL preparation in Taiwan.

In the Arabic and Spanish samples, there was a slight trend for the graduates to score higher than the undergraduates, but this trend is more remarkable for how small it is than for the few correlations that are statistically significant.

Although there were a few significant correlations between major field designations and some of the criterion scores, there was no evidence that the writing sample or the TOEFL was more sensitive to major field differences. Thus, except for the differential sensitivity of the writing sample for the Chinese group noted above, the available evidence suggests that the writing sample and the TOEFL are comparably sensitive to differences in age, sex, undergraduate-graduate status, graduate major, and number of years of studying English.

Correlational Analyses

The correlations between the scores on the various measures provide additional information regarding the validity of the TOEFL and GRE General Test scores for this sample of candidates. This section reports the correlations between TOEFL scores and direct measures of writing (writing samples), correlations between GRE General Test scores and direct and indirect measures of writing, and correlations between the scores on the direct measures obtained by the different scoring methods (holistic, discourse/sentence, and subject matter). The final section describes the data obtained on the Writer's Workbench.

Correlations with TOEFL Scores

Intercorrelations of the various writing sample scores with TOEFL scores are presented in Table 16. Consistent with the previous discussion of the lack of differentiation between holistic scores and discourse/sentence scores and between the two topic types, the correlations of each writing sample score with a given TOEFL score were essentially identical. Because it is the most reliable score, the total holistic score correlated most highly with the TOEFL. The correlations of .72 between the holistic total and the TOEFL total indicates that the two measures are largely overlapping, but that the overlap is not perfect. Because of the high reliabilities of both the writing sample holistic total (about .90) and the TOEFL total (about .95), correcting the correlation for attenuation does not substantially alter the conclusion (corrected correlation of TOEFL and holistic total =.78). The writing sample is measuring some component of English proficiency that is not assessed by the TOEFL. To better understand the degree of overlap or independence, note that the correlation between the holistic total and TOEFL structure and written expression (.69) is just about the same as the correlation of TOEFL listening comprehension and TOEFL structure and written expression (.68). Thus, if the writing sample were a fourth section of the TOEFL, its relationship with the other measures would be consistent with the degree of relationship among sections observed in the present test. It is important to note that the writing sample measures some higher order organizational skills that even native speakers may find difficult. Thus, the writing sample should be expected to tap some skills that are well beyond the minimal proficiency level emphasized in the TOEFL.

Although the discourse and sentence scores were initially conceived as two separate scores that could not be summed, the high correlation between them suggested that a sum score, with its increased variance, might correlate more highly with the TOEFL total. Therefore, an additional score was created by adding the discourse total (sum of the discourse scores over two raters on each of the two topics, Space and Farming) to the sentence total. This discourse/sentence total correlated .73 with the TOEFL total and was nearly identical to the correlation of the holistic total with the TOEFL total (.72). But the discourse/sentence total was based on scores from only two essays; a holistic total based on holistic scores from only those two essays correlated .68 with the TOEFL total. Thus, the discourse/sentence total appears to yield slightly better predictions. Additional research is needed to fully explain this apparent advantage. It may result simply from the increased variance of the discourse/sentence total; using an expanded score scale for the holistic judgments might make both scores more comparable. An additional consideration is the possibility of training effects. In this study all discourse/sentence scores were assigned on the second day of a two-day scoring session with holistic scores having been assigned on the first day. Judges were therefore more familiar with the range of student responses during the discourse/sentence scoring than they had been during the holistic scoring. Future research should counterbalance the order effect or, preferably, use totally different groups of judges for the two kinds of scoring.

Correlations with GRE General Test Scores

Intercorrelations of the various scores on the direct measures of writing ability (writing samples) and indirect measure of writing ability (LSAT writing test) appear in Table 22. For the foreign sample, the correlation of scores on the writing sample with TOEFL total scores and GRE verbal scores are nearly identical; scores on the writing sample could be predicted equally well from GRE verbal or TOEFL scores. The correlation of writing sample scores with GRE verbal scores is substantially higher in the total sample than in the foreign sample because the United States students scored relatively high on both measures. The moderately high correlations of all writing sample scores with scores on GRE analytical suggest the contribution made by English (verbal) proficiency to the analytical section; however, this correlation, when compared to the correlation of the writing sample scores with GRE verbal, also suggests that this section assesses an ability other than one that is purely verbal. The low negative correlations of the writing sample scores with GRE quantitative scores reflect a pattern that further reinforces the independence of quantitative scores from verbal and analytical scores. These relationships are reinforced by the correlations of GRE verbal scores with GRE analytical scores (.62) and with GRE quantitative (-.17) scores, as well as by the correlations of GRE analytical scores with GRE quantitative scores (.33). Because this sample of data consists preponderantly of foreign students, the GRE General Test scores present remarkably stable patterns of relationships. It should be noted that the negative correlations with the GRE quantitative score are an artifact of a foreign sample in which candidates with very low GRE verbal scores may still seek admission if their GRE quantitative scores are very high. In the sample of United States students, GRE verbal and GRE quantitative were correlated .64.

For the foreign student sample (N = 124), the correlations of scores on the TOEFL with scores on the GRE General Test, although slightly lower, repeat the same general patterns. GRE verbal scores are more highly correlated with TOEFL scores, particularly with Section III (Reading Comprehension) TOEFL scores (.72) and with the total TOEFL score. These correlations support the relationship of the GRE verbal to the TOEFL as measures of English language proficiency. Since the GRE verbal items place emphasis on reading comprehension, the correlation with Section III of the TOEFL provides further evidence that reading comprehension contributes substantially to the verbal scores. The low correlations of TOEFL scores with GRE quantitative scores repeat the pattern observed previously, as do the moderate correlations of TOEFL scores with the GRE analytical scores. These TOEFL/GRE correlations, when compared with the correlations of the three sections of the TOEFL, clearly indicate that the TOEFL assesses English proficiency overall, but that each section of the TOEFL contributes a somewhat different measure of that proficiency.

Finally, for the United States student sample (N=43,) the scores on the indirect measure of writing ability present some interesting patterns. The high correlations are those of the GRE verbal scores with the sections of the LSAT writing test, whereas the lowest correlation is that with scores on the usage section and the GRE quantitative score. The correlations of the quantitative and analytical sections of the GRE with the scores on the sentence correction section and the total score on this indirect measure are low. The correlations of the scores on the writing samples with the LSAT writing test are not as high as would be expected, the highest being the correlation between scores on the sentence correction section of the LSAT writing test and the total holistic scores (only .51). However, correlations may have been attentuated by the ceiling-level writing sample performance of many of the United States students. Although the students in the United States sample performed very well on the writing samples, their scores on the indirect measure of writing ability did not reflect the same degree of "writing competence." The high correlation of scores on GRE verbal, an indirect measure, with one section of the indirect measure of writing ability, when compared to the correlations of writing sample and LSAT writing test scores, suggests that the method of assessment (direct vs. indirect) may elicit different levels of performance. This difference in performance on direct and indirect measures of writing ability, although they may assess some overlapping abilities, was indicated in the two-factor solution to the TOEFL and writing samples scores discussed in a previous section of this chapter. The score differences reflect level differences, but also may be influenced by the kinds of writing abilities that are elicited by the direct and indirect measures, for which further research would be required.

An additional analysis focused on the correlations of writing sample scores with item types, or item parcels, in the GRE General Test. The sample of subjects was reduced to restrict the analysis to the three forms of the GRE that were taken by most of the candidates in our sample, thus eliminating small numbers of subjects who took other forms of the test. This sample of 132 subjects consisted of 21 candidates for whom English is their primary language, 5 Arabic-language candidates, 73 Chinese-language candidates, and 33 Spanish-language candidates. The GRE score data was retrieved, and separate scores were obtained for the different item types that make up the test. These scores were correlated with the total holistic score, averaged over four writing samples, on the direct measures of writing ability (Table 23). The observed pattern of correlations was consistent with the relationships reported in other GRE studies. Specifically, the analytical reasoning and logical reasoning scores were not highly correlated (.24), and the analytical reasoning items were more highly correlated with the quantitative items (.46, .35, .50) than were the logical reasoning items (-.09, -.18, .02). On the other hand, the logical reasoning items were more highly correlated with the verbal items (.65, .50 .67) than were the analytical reasoning items (.15, .17, .24). The holistic scores were more highly correlated (.64) with the logical reasoning items and with the three types of verbal items (.68, .67, .70) than with the analytical reasoning items (.23). This result indicates that the holistic scores, as expected, reflect verbal ability, as also is reflected in the logical reasoning items.

Table 24 reports the results of a stepwise regression analysis of these data, which parallel the correlational analysis. The prediction of the total writing sample score is enhanced somewhat by the addition of scores on the two types of verbal items (reading comprehension and the discretes--antonyms and analogies) and next by scores on logical reasoning items and verbal sentence completion items. The quantitative item types, as well as the analytical reasoning items, do not contribute substantially to the holistic score.

Writer's Workbench Analyses

The Writer's Workbench, in addition to serving as a tool for editing and instruction, appears to have promise as a research tool. The relationships of features of writing identified on the Workbench with other approaches to evaluating the features of a writing sample (e.g., holistic scores, error analyses) provide somewhat detailed evidence about these features. The data analyzed on the Workbench for this study suggest that certain characteristics of writing that are attended to by a human reader are related to, and therefore are likely to have influenced, the evaluation of a piece of writing. These data provide some interesting clues, which need to be investigated with further research. The results provide additional information about the features of writing that readers may not be conscious of, but that may contribute to a score. This observation is parallel to the experience of the readers who sensed that they were attending to somewhat different features of writing when applying the discourse/sentence method of scoring and would have expected the discourse/sentence and holistic scoring methods to yield different scores.

Tables 25 through 29 summarize the correlational relationships between the various Workbench features and TOEFL and writing sample scores. The data consist of four sets of writing samples that were selected to be representative on the basis of the range of holistic scores assigned to the different subsamples of the total sample--Arabic, Chinese, Spanish, English, graduate, undergraduate, hard science, and social science. These small samples appear to be representative, in that they reflect the same relationships that were observed for the total sample with respect to scores on the writing samples and sections of the TOEFL. The correlations of the data on the Workbench features show one pattern that confirms the validity of the TOEFL and the writing sample scores as English language proficiency measures--the highest correlations obtained are with the TOEFL and the holistic and discourse/sentence scores (only the Farming and Space topics received discourse/sentence scores). The scores on the writing samples yielded additional, relatively high correlations that were not found with the TOEFL. Characteristics such as "number of words," "number of content words," "number of short sentences," "number of 'to be' verbs" (comparing Tables 25 through 29) have moderate correlations with the writing sample scores.

Scores on papers written on the different topics also yielded significant, though moderate correlations with somewhat different Workbench variables. For example, some significant correlations of holistic scores with Writer's Workbench features obtained for three of the topics were not observed for the Continents topic (Table 28)--"number of short sentences, " "number of long sentences," "percentage of passives," "number of content words," "percentage of prepositions" and "percentage of conjunctions." With a larger sample of papers, it would be worthwhile to investigate whether patterns of correlations of Workbench scores with writing sample scores differ in significant ways for papers written on different topics and in different discourse modes. Since the readers exhibited high agreement across topics and topic types, the potential finding that differential features contributed to the same numerical ratings would be of interest, because it would suggest that readers are able to adjust their standards to account for different features of writing elicited by different topics or modes of discourse.

These correlations should be viewed only as descriptions of the relationships observed within four discrete sets of data, however. Since a large number of variables were correlated and a small number of representative samples of papers were analyzed for each of the four topics, these correlations may have resulted largely due to chance factors. Before any inferences or conclusions can be drawn, this study requires replication with larger samples of papers and a possible reduction in the number of variables. We are reporting these data because they suggest some interesting relationships among features of the papers and the scores assigned to the papers, relationships that warrant additional exploration. Thus the results can be regarded as descriptive of only these particular sets of data--papers written in response to four specific topics, subjected to specific scoring procedures and systems, and within the context of the parameters of this research study.

None of the individual features analyzed by the Writer's Workbench is highly correlated with TOEFL section scores or writing sample scores (holistic or discourse/sentence); therefore a specific Workbench feature would not serve as a predictor of TOEFL section scores or of writing sample scores. Instead, the separate features of papers obtained from the Workbench system tend to support the notion that several different features contribute to the quality of a writing sample. Tables 30 through 32 report stepwise regression analyses conducted on the Writer's Workbench, Section II of the TOEFL (structure and written expression), and writing sample score variables for these sets of data.

For these analyses, we reduced the number of Writer's Workbench features by eliminating features that introduced redundancy because they were highly correlated. For example, "percentage of short sentences" was dropped, but "number of short sentences" was retained. Tables 30 through 32 list, for each topic, the Writer's Workbench features that would contribute to the prediction of TOEFL Section II scores (Table 30), the holistic scores (Table 31), and the discourse/sentence scores (Table 32). Some Writer's Workbench features, which represent the features of the writing samples in each set of papers, appear to contribute consistently to the prediction of any of the scores--features such as "number of content words" and "number of spelling errors." Other, somewhat different features contributed to papers written in response to different topics--features such as "number of long sentences" for the holistic scores for the Farming topic. Thus these analyses provide a rough approximation of the most important Writer's Workbench correlates with Section II of the TOEFL and the writing sample score variables.

The data should be interpreted cautiously because the Writer's Workbench also is not infallible--it is capable only of doing counts and calculations based on the tangible characteristics of a paper (e.g., word counts, readability formulas). Occasionally, it is not totally accurate, such as identifying a spelling error when the word has been correctly spelled. In such instances, we did not accept the output at face value. The spelling errors, printed on the output, were carefully checked, and correctly spelled words were not tallied. However, some of the internal judgments made by the programs that are not printed cannot be checked. With a recognition of its limitations, the Writer's Workbench probably can be considered more reliable than a human judge, particularly in cases where the features are objectively identifiable and can be counted accurately by a computer program. The CSU version of the Workbench offers judgmental comments to the writer regarding features of a paper that generally are not considered "good" writing. One example is the overuse of "to be" verbs. For our purposes, the counts of "to be" verbs provide objective data, without attaching judgments. In fact the CSU staff noted that the chart/graph topics seemed to elicit more "to be" verb usage, which was appropriate because other verbs are not as likely to be used in clearly describing a chart or graph.

VI. SUMMARY OF RESULTS AND CONCLUSIONS

This research generated a considerable amount of information contributing to the validity of measures of English language proficiency--writing samples, the TOEFL, and the GRE General Test. A summary of the major findings follows:

- The two scoring methods for the writing samples, holistic and discourse-level/sentence-level (D/S), yielded essentially the same mean levels of performance and were highly correlated, indicating that the two-score method may not provide any significant advantage over the one-score method. Aside from the high correlations among holistic and discourse/sentence scores, we observed that (1) it was very difficult to select sample papers for scoring sessions that represented reliably different values of D and S, and (2) although readers could agree on the levels of performance for D and S, they perceived the constructs of discourse-level and sentence-level features to be unclear and confounded (thus challenging the validity of separating judgments on the basis of D and S).

- The means of the writing sample scores reflected level differences for the three language groups for whom English is not their primary language. For every writing sample score, the means were lowest for the Arabic sample, in the middle for the Chinese sample, and highest for the Spanish sample.

- The mean holistic and discourse/sentence scores obtained by the sample of United States candidates on the writing samples were considerably higher than the mean scores for the foreign group, not a surprising result since the focus of the study was on measures that assess English language proficiency.

- The reliabilities of all the scores assigned to the writing samples were remarkably high, indicating that the consistent scoring of writing samples can be achieved (under the optimal scoring conditions described in previous chapters). The various types of evidence for reliability of the holistic scores consisted of interrater reliability, reliability across topics, and reliability within language groups. For the discourse-level and sentence-level scores, evidence for reliability consisted of interrater reliability, reliability across score types and across topics, reliability within language groups, and reliability across ESL and English readers.

- Correlations were as high across topic type as within topic type. This result suggests that (1) the different topics did not elicit qualitatively different writing performance, and/or (2) the readers maintained a comparable scale for evaluating the writing samples, despite performance fluctuations from topic to topic.

These positive results, however, should not be interpreted as evidence that papers written in response to any topic or type of topic would yield equivalent reliability. The topics were selected on the basis of previous research indicating that specific kinds of topics would serve as more appropriate stimuli to reflect the academic writing task demands experienced by students in higher education in the United States. Carefully controlled conditions of design and pretesting, and of scoring methods that emphasized functional academic English proficiency, would need to be replicated to attain similar results.

Both this study and our previous survey of academic writing tasks have demonstrated, though, that topics designed to elicit the English skills of TOEFL candidates in different disciplines do not need to be subject-specific in order to evaluate writing performance effectively as long as they are within the context of relevant academic competencies.

- Whatever differences in the perception of good writing may exist among regular English teachers, ESL teachers, social science teachers, and engineering teachers, these differences do not interfere with the ability of these diverse groups to rank students' writing samples in the same order. When subject-matter experts in engineering and the social sciences were asked to rate representative subsamples of papers written in response to two topics, the professors' ratings were highly correlated with each other--the mean social science ratings correlated .92 with the mean engineering ratings for each of the two topics. When compared with the holistic scores assigned during the regular scoring session for the compare/contrast topic (Space), the mean social science judgment correlated .86 with the holistic scores, and the mean engineering judgment, .92. For the chart/graph topic (Farming), the correlations were .83 and .82, respectively. This outcome further supports the assumption that general agreement exists, even when not formally identified and verbalized, concerning standards for academic writing competence.

  These results also can be explained by two design factors: (1) the professors were instructed to evaluate the papers from the perspective of writing competence required of students to succeed in their graduate-level departments, as opposed to writing competence in general; and (2) they were supplied with a limited number and representative sample of papers such that the task was to some extent more highly structured than the task addressed by the holistic readers.

- The reader responses to the questionnaires provided information about the points of view with regard to the evaluation of writing skills and the readers' exposure to different methods of scoring papers on the same topics. Reader ratings of the features of written assignments suggested that the readers perceived that they were attending to somewhat different characteristics of writing competence during the holistic scoring than during the discourse/sentence scoring. However, although the readers may have focused on different features, the means and standard deviations of the scores indicated that the different scoring methods did not yield different score levels. Thus the evaluations of the quality of writing competence were consistent, regardless of scoring method. These results suggest that papers that are strong on one measure (D) are strong on another (S), or that perceptions of D and S go hand in hand. This finding also supports the supposition held by readers of compositions that general agreement exists, even when not formally identified and verbalized, concerning the standards for writing competence.

Data obtained from the Writer's Workbench, as a tool for investigating the features of writing samples that may be salient to readers, suggested that further investigation may provide useful information regarding relationships among features of the papers and the scores assigned to the papers.

- In response to other questions on the questionnaire, a considerable number of readers (70 percent) felt that the scores they were asked to assign during both scoring sessions were appropriate to the particular sample of papers.

- Many readers indicated that they would be very uncomfortable attempting to assign descriptions to score levels because individual papers at one score level can differ considerably. Most readers appeared to agree, however, that sample papers at each score level could be useful and meaningful if provided in a score manual for an operational writing sample testing program, both to other readers of writing samples and to those who would interpret writing sample scores.

- A principal axes factor analysis with varimax rotations of holistic scores and TOEFL section scores resulted in a two-factor solution. The two factors appear to be method factors, one consisting of scores on the three sections of the TOEFL and the other, of holistic scores on papers written in response to the four topics. One interpretation of the two factors suggests that performance on measures of English language proficiency becomes more differentiated when the measures require a candidate to respond by applying different cognitive processes--recognition vs. production.

- A comparison of the relationships of writing sample and TOEFL mean scores showed that the pattern of means across the three language groups is highly consistent. This lack of interaction between type of score (writing sample or multiple-choice) and language group is consistent with the notion that both types of scores may assess, to a great extent, the same underlying language proficiency dimension. However, there is some evidence that the between-groups differences are smaller for the scores on the writing samples than for TOEFL scores.

- The correlations between the holistic score total (direct evidence of a productive skill) and the TOEFL total score (measures of receptive skills and indirect measures of writing) indicate that the two measures evaluate English proficiency to a considerable degree, but that the overlap between the two instruments is not perfect. The writing sample contributes additional information regarding English proficiency, since a competently executed writing sample demonstrates the application of cognitive abilities far beyond the mastery of mechanics. The TOEFL provides evidence of mastery of English language skills, but not of higher-order writing skills such as organization and quality of ideas.

  In addition, the relationships of the writing sample score with other sections of the TOEFL are consistent with the pattern of relationships among the TOEFL sections, such as reported in previous research (Pitcher & Ra, 1967; Pike, 1979), although the sizes of the correlations obtained in this study are somewhat lower. The earlier research results, however, cannot be compared directly with our findings because of basic design differences. In the previous studies, the composition of the TOEFL was different, since it was the five-section version used prior to 1976. In addition, the topics differed considerably—topics in Pike's study included more explicit and restrictive instructions and elicited papers written in a narrative form. Pike also investigated three native country groups (from Chile, Peru, and Japan) whereas this research targeted a different configuration of native languages (Arabic, Chinese, Spanish). The consistent pattern of relationships obtained in the three studies, however, lend further support to the validity of the TOEFL and direct measures of writing ability.

- For the foreign sample, the correlation of scores on the writing sample with the TOEFL total scores and with GRE verbal scores is nearly identical, indicating that the writing sample scores serve as an indicator of English language skills. For foreign candidates, however, the GRE verbal measure requires a high level of English proficiency in contrast to the TOEFL.

- The correlation of writing sample scores with GRE verbal scores is substantially higher in the total sample than in the international sample because the United States students scored relatively high on both measures. The correlations of scores on sections of the GRE General Test with the TOEFL and writing sample scores present remarkably stable patterns of relationships.

- When the holistic writing sample scores, averaged over four topics, were related to scores on item types within the sections of the GRE General Test, the observed pattern of correlations was consistent with the relationships reported in other GRE studies. Specifically, the analytical reasoning and logical reasoning items were not highly correlated, and the analytical reasoning items were more highly correlated with the quantitative items than were the logical reasoning items. On the other hand, the logical reasoning items were more highly correlated with the verbal items than were the analytical reasoning items. The holistic scores were more highly correlated with the logical reasoning items than with the analytical reasoning items, further indication that the holistic scores reflect verbal ability as measured by relevant item types in the GRE General Test.

Conclusions

The results suggest that, with careful topic selection and adequate training of raters, writing samples can provide a reliable measure of the English proficiency of nonnative speakers as well as native speakers of English, and that direct measures of writing performance, although substantially correlated with multiple-choice measures such as the TOEFL and GRE General Test, contribute additional information regarding English proficiency.

There was no indication of any important differences between the two topic types (chart/graph interpretation and compare/contrast) used in this study. However, it is important to remember that both topic types represent structured, academically oriented writing; results may have been different with a "What I did last summer" type of topic. Furthermore, although a single topic type might be all that is needed in an operational program, that does not imply that a single topic is sufficient. Different topics, even within the same topic type, elicit slightly different performances, and the reliability of the total score increases as the number of topics sampled increases.

Separate scores for discourse-level and sentence-level skills do not appear to present any advantage over a single holistic score. Computer scoring of writing samples (Writer's Workbench) provides data that appear to be potentially useful for assisting writing instruction and in the development of scoring systems, but it is not a substitute for holistic scoring based on human judgments.

Writing performance clearly differs across language groups, just as TOEFL performance differs across language groups. But there is no evidence that the writing samples unfairly discriminate against any group. Again, careful topic selection procedures must be emphasized. Some of the topics rejected during the pilot testing did indeed appear to be discriminatory. Further research with criterion scores that were independent of TOEFL scores would be needed to fully answer any questions of possible bias.

Recommendations

From the standpoint of the TOEFL program, this research contributes valuable information regarding the potential addition of direct measures of academic writing ability to the TOEFL. Based on our findings, we recommend that the decision making regarding this issue take into account the following considerations:

1. A program of topic design and development such as that used in this study, and involving pretesting to investigate the efficacy of new topics and of relationships among performances on new topics with sections of the TOEFL, should be implemented. The latter objective could be met by including topics for pretesting during actual TOEFL administrations at selected international sites.

   If a score for the direct assessment of writing becomes operational in a testing program, eventually we would expect to observe changes in the size and, possibly, patterns of correlations with other sections of the test. The inclusion of a writing sample communicates a message about what is valued in the assessment of English proficiency. Institutions that prepare foreign candidates for admission to postsecondary institutions in the United States, as well as the candidates themselves, undoubtedly will take steps to meet the challenge of a direct measure of writing ability, resulting in observed changes in performance on that measure and on other measures of English proficiency.

2. Additional validation research that relates performance on writing samples to writing performance in academic settings should be conducted.

3. If direct measures of writing are implemented, one writing task is not necessarily a sufficient sample of writing performance, since it would not provide assurance that a candidate would perform consistently on other writing tasks. The results of this study could be interpreted to suggest that performance on one writing assignment provided valid and reliable information regarding performance on the other tasks; with new topics, a different (possibly more or less heterogeneous) population, and under slightly different testing conditions, however, this finding would need to be demonstrated. Initially, a new operational program should determine that performance and the evaluation of that performance are consistent from one writing assignment to another. Ideally, each candidate should be required to respond to more than one writing assignment in the early stages of a direct assessment program.

4. The number of scorers who evaluate a paper present a significant operational cost consideration. At least two readers should be used to ensure valid and reliable scores, particularly when those scores may be critical to the educational progress of candidates. It may be possible that, after accumulating a history of highly correlated scores assigned by two readers, the program could justify scoring by only one reader.

5. Meaningful information regarding the appropriate use and intepretation of scores on direct measures of writing should be provided to those who interpret and use these scores. The consensus of the readers involved in the holistic and discourse/sentence scoring sessions indicated that the different points on a score scale cannot be described in terms of the salient features of the papers at each level, since so many different features are mentally weighed in the course of making a holistic judgment, and these features vary from paper to paper at a particular score point. Instead, they recommended a manual that contains several benchmark papers at each point along the score scale, with descriptive comments accompanying each paper. Such information would assist test users in making placement decisions that would be appropriate to the candidate and to the institution's specific academic requirements.

These general recommendations represent a variety of specific operational issues that will need to be resolved and that do not fall within the domain of this study, particularly the criteria for making the decision whether or not to include a direct measure of writing ability as a section of the TOEFL. Based on the results of this research, either decision could be justified.

From the standpoint of the GRE program, the data have contributed valuable information regarding the relationships among GRE General Test scores, TOEFL scores, and direct measures of writing ability. These data contribute to the interpetation of GRE score data, since considerable numbers of GRE candidates are nonnative speakers of English, and writing ability is important to the successful performance of both native and nonnative speakers in graduate-level academic contexts.

Bibliography

American Psychological Association. (1984, February). Joint technical standards for educational and psychological testing, 4th Draft. Washington, DC: American Psychological Association, Office of Scientific Affairs.

Angelis, P. J. (1982). Language skills in academic study. Unpublished manuscript. Princeton, NJ: Educational Testing Service.

Blanton, L. L. (1982). ESP: Benefits for all of ESL. English for Special Purposes, 64, 6-7.

Boyan, D. R. (Ed.). (1981). Open doors: 1980/81 report on international education exchange. New York: Institute of International Education.

Breland, H. M., & Jones, R. J. (1982). Perceptions of writing skill. (ETS Research Rep. No. 1982-47). Princeton, NJ: Educational Testing Service.

Bridgeman, B., & Carlson, S. (1983). Survey of academic writing tasks required of graduate and undergraduate foreign students. (ETS Research Rep. No. 1983-18). Princeton, NJ: Educational Testing Service.

Buckingham, T. (1979). The goals of advanced composition instruction. TESOL Quarterly, 13, 241-254.

Canale, M. (1983). On some dimensions of language proficiency. Chapter 20. In J. W. Oller, Jr. (Ed.), Issues in language testing research. Rowley, MA: Newbury House Publishers.

Canale, M., & Swain, M. (1979). Communicative approaches to second language teaching and testing. Ontario: Ontario Ministry of Education.

Carlson, S., & Bridgeman, B. Testing ESL student writers. In K. Greenberg, H. Wiener, & R. Donovan (Eds.), Writing assessment: Issues and strategies. New York: Longman (in press).

Carpenter, C., & Hunter, J. (1981). Functional exercises: Improving overall coherence in ESL writing. TESOL Quarterly, 15, 425-434.

Carrell, P. L. (1982). Cohesion is not coherence. TESOL Quarterly, 16(4), 479-488.

Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In Testing the English proficiency of foreign students. Washington, DC: Center for Applied Linguistics. Reprinted in H. B. Allen and R. N. Campbell (Eds.), Teaching English as a second language: A book of readings. New York: McGraw-Hill.

Cherry, L. L., Fox, M. L., Frase, L. T., Gingrich, P. S., Keenan, S. A., & Macdonald, N. H. (1983). Computer aids for text analysis. Bell Laboratories Record, May/June.

Clark, J. L. D. (1972). Foreign language testing: Theory and practice. Philadelphia: Center for Curriculum Development.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement. Washington, DC: American Council on Education, pp. 443-507.

Cummins, J. (1983). Language proficiency and academic achievement. In J. W. Oller, Jr. (Ed.), Issues in language testing research. Rowley, MA: Newbury House Publishers.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). Factors in judgments of writing ability. (ETS Research Bull. 1961-15). Princeton, NJ: Educational Testing Service.

Dubin, F., & Olshtain, E. (1980) The interface of writing and reading. TESOL Quarterly, 14(3), 353-363.

Eskey, D. E. (1983). Meanwhile, back in the real world...: Accuracy and fluency in second language teaching. TESOL Quarterly, 17(2), 315-323.

Farr, M. (1983, January). What research is contributing to writing assessment. In D. A. McQuade & V. B. Slaughter (Eds.), Notes from the National Testing Network in Writing. New York: Instructional Resource Center, City University of New York, p. 19.

Freedman, S. W. (1979). How characteristics of student essays influence teachers' evaluations. Journal of Educational Psychology, 71, 328-338.

French, J. W. (1962). Schools of thought in judging excellence of English themes. Princeton, NJ: Educational Testing Service.

Godshalk, F. I., Swineford F., & Coffman, W. E. (1966). The measurement of writing ability. New York: College Entrance Examination Board.

GRE guide to the use of the Graduate Record Examinations program, 1983-1984 (1983). Princeton, NJ: Educational Testing Service.

Greenberg, K. L. (1983). Writing tasks and students' writing performance (1). In B. Kwalick, M. Silver, & V. B. Slaughter (Eds.), Selected Papers from the 1982 Conference 'New York Writes'." New York: Instructional Resource Center, The City University of New York.

Hale, G. A. & Hinofotis, F. B. (1981, 1983). New directions in English language testing. Princeton, NJ: Educational Testing Service.

Hill, S. S., Soppelsa, B. F., & West, G. K. (1982). Teaching ESL students to read and write experimental-research papers. *TESOL Quarterly*, *16*, 333-347.

Hunt, K. W. (1970). Syntactic maturity in school children and adults. *Monographs of the Society for Research in Child Development*, *53*(1), (Serial No. 134).

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House Publishers.

Johns, A. M. (1980). Cohesion in written business discourse: Some contrasts. *The ESP Journal*, *1*(1), (Fall), 35-44.

Johns, A. M. (1981). Necessary English: A faculty survey. *TESOL Quarterly*, *15*(1), 51-57.

Jones, R. L. (1977). Testing: A vital connection. In J. K. Phillips (Ed.), *The language connection: From the classroom to the world*. Skokie, IL: National Textbook Co.

Kaplan, R. B. (1966). Cultural thought patterns in inter-cultural education. *Language Learning*, *16*, 1-20.

Kaplan, R. B. (1972). The anatomy of rhetoric: Prolegomena to a functional theory of rhetoric. In *Language and the teacher: A series in applied linguistics, 8*. Philadelphia, PA.: Center for Curriculum Development.

Kaplan, R. B. (1976). A further note on contrastive rhetoric. *Communication Quarterly*, *14*(2), 12-19.

Kaplan, R. B. (1977). Contrastive rhetoric: Some hypotheses. *ITL*, *39-40*, 61-72.

Kaplan, R. B. (1982). Contrastive rhetoric: Some implications for the writing process. In I. Pringle, A. Freedman, & J. Yalden (Eds.), *Learning to write: First language, second language*. London: Longman.

Keech, C. (1982, November). Designing prompts for holistic writing assessments: Knowledge from theory, research, and practice. Part II. Practices in designing writing test prompts: Analysis and recommendations. In L. Ruth (Project Director), *Properties of writing tasks: A study of alternative procedures for holistic writing assessment*. (Final Report NIE-G-80-0034). Berkeley, CA: Bay Area Writing Project, Graduate School of Education, University of California.

Kiefer, K. E., & Smith, C. R. (1984). Textual analysis with computers: Tests of Bell Laboratories' computer software. Research in the Teaching of English, 17(3), 201-214.

Kroll, B. (1979). A survey of the writing needs of foreign and American college freshmen. English Language Teaching Journal, 33, 219-226.

Kuhn, T. (1970). The structure of scientific revolutions (2nd ed.). Chicago: University of Chicago Press.

Larsen-Freeman, D. (1978). An ESL index of development. Paper read at the Twelfth Annual TESOL Convention, Mexico (D.F.), April 4-9.

Larson, R. L. (1983, January). Tests of writing ability: Their messages for students and teachers. In D. A. McQuade & V. B. Slaughter (Eds.), Notes from the National Testing Network in Writing. New York: Instructional Resource Center, City University of New York, pp. 8-9.

Lay, N. D. S. (1982). Composing processes of adult ESL learners: A case study. TESOL Quarterly, 16, 406.

Lindstrom, M. W. (1981). Native speaker reactions to stylistic errors in writing: An error evaluation. Unpublished master's thesis, Colorado State University, Fort Collins, CO.

Lloyd-Jones, R. (1982, October). Skepticism about test scores. In K. L. Greenberg, H. S. Wiener & R. A. Donovan (Eds.), Notes from the National Testing Network in Writing. New York: Instructional Resource Center, City University of New York, pp. 3, 9.

Macdonald, N. H., Frase, L. T., Gingrich, P. S., & Keenan, S. A. (1982). The Writer's Workbench: Computer aids for text analysis. Educational Psychologist, 17(3), 172-179.

Moffett, J. (1968). Teaching the universe of discourse. Boston: Houghton Mifflin.

Morrow, K. E. (1977). Techniques of evaluation for a notional syllabus. Reading, England: Centre for Applied Language Studies, University of Reading (Study commissioned by the Royal Society of Arts).

Munby, J. (1978). Communicative syllabus design. Cambridge: Cambridge University Press.

Navy Personnel Research and Development Center. (1982). Readability formulas: Their application in the armed forces (NPRDC SR 83-8). San Diego.

Odell, L. (1981). Defining and assessing competence in writing. In C. R. Cooper (Ed.), The nature and measurement of competency in English. Urbana, IL: National Council of Teachers of English, pp. 95-138.

Oller, J. W., Jr. (1979). Explaining the reliable variance in tests: The validation problem. In E. Briere & F. Hinofotis (Eds.), Concepts in language testing: Some recent studies. Washington, DC: TESOL, pp. 61-74.

Ostler, S. E. (1981). A survey of academic skills in the low-level ESL class. TESOL Quarterly, 14, 489-502.

Pearson, C. R. (1981, December). Advanced academic skills in the low-level ESL class. TESOL Quarterly, 15, 413-423.

Pike, L. W. (1979). An evaluation of alternative item formats for testing English as a foreign language (TOEFL Research Report No. 2). Princeton, NJ: Educational Testing Service.

Pitcher, B., & Ra, J. B. (1967). The relationships between scores on the Test of English as a Foreign Language and the ratings of actual theme writing (Statistical Rep. No. 67-9). Princeton, NJ: Educational Testing Service.

Purves, A. (1984, March 8). International perspectives on writing assessment. Papers presented at the National Testing Network in Writing: Second Annual Conference on Writing Assessment, Tallahassee, FL.

Quellmalz, E. S., Capell, F. J., & Chou, C. (1982). Effects of discourse mode and response mode on the measurement of writing competence. Journal of Educational Measurement, 19, 241-258.

Quintanilla, R. (1983). Articulating ESL instruction in New York City high schools and CUNY. In B. Kwalick, M. Silver, & V. B. Slaughter (Eds.), Selected Papers from the 1982 Conference 'New York Writes'. New York: Instructional Resource Center, The City University of New York.

Readability formulas: Their application in the armed forces (1982, November) (NPRDC SR 83-8). San Diego: Navy Personnel Research and Development Center.

Richards, J. (1979). Rhetorical and communicative styles in the new varieties of English. Language Learning, 29(1), 1-26.

Ruth, L. (1982, November). Designing prompts for holistic writing assessments: Knowledge from theory, research, and practice. Part I: Sources of knowledge for designing writing test prompts. In L. Ruth (Project Director), Properties of writing tasks: A study of alternative procedures for holistic writing assessment. (Final Report NIE-G-80-0034). Berkeley, CA: Bay Area Writing Project, Graduate School of Education, University of California.

Selinker, L., Todd-Trimble, M., & Trimble, L. (1978). Rhetorical function-shifts in EST discourse. TESOL Quarterly, 12(3), 311-320.

Shaughnessy, M. P. (1977). Errors and expectations. New York: Oxford University Press.

Smith, C. R., & Kiefer, K. (1982, April). Writer's Workbench: Computers and writing instruction. Paper presented at the Proceedings of the Future of Literacy Conference, University of Maryland, Baltimore, MD.

Spack, R., & Sadow, C. (1983). Student-teacher working journals in ESL freshmen composition. TESOL Quarterly, 17(4), 575-593.

Swineford, F. (1964). Test analysis, Advanced Placement Examination in American history, form MBP. (Statistical Report No. 1964-53). Princeton, NJ: Educational Testing Service.

Takala, S., Purves, A. C., & Buckmaster, A. (1982). On the interrelationships between language, perception, thought and culture and their relevance to the assessment of written composition. In B. H. Choppin & T. N. Postlethwaite (Eds.), A. C. Purves & S. Takala (Guest Eds.), Evaluation in education: An international review series. An international perspective on the evaluation of written composition. New York: Pergamon Press, pp. 317-342.

Taylor, B. P. (1982). Content and written form: A two-way street. TESOL Quarterly, 15(1), 5-13.

Taylor, B. P. (1982, May). Teaching ESL: A communicative, student-centered approach. "Student initiative in the ESL class." Paper presented at the 16th Annual ESL Convention in Honolulu, Hawaii.

Thompson-Panos, K., & Thomas-Ruzic, M. (1983). The least you should know about Arabic: Implications for the ESL writing instructor. TESOL Quarterly, 17 (4), 609-623.

Test of English as a Foreign Language. (1983). TOEFL test and score manual. Princeton, NJ: Educational Testing Service.

Troyka, L. Q. (1982, October). Looking back and moving forward. In K. L. Greenberg, H. W. Wiener, & R. A. Donovan (Eds.), Notes from the National Testing Network in Writing. New York: Instructional Resource Center, City University of New York, pp. 3, 9.

Walz, J. C. (1982). Error correction techniques for the foreign language classroom. Washington, DC: Center for Applied Linguistics.

Weaver, B. T. (1982, October). Competency testing and writing program development. In K. L. Greenberg, H. S. Wiener & R. A. Donovan (Eds.), Notes from the National Testing Network in Writing. New York: Instructional Resource Center, City University of New York, October 1982, p. 13.

West, G. K. (1982). Engineering faculty evaluations of foreign graduate student writing. Gainesville, FL: University of Florida. Unpublished manuscript.

West, G. K., & Byrd, P. (1982). Technical writing required of graduate engineering students. Journal of Technical Writing and Communication, 12, 1-6.

Widdowson, H. G. (1974). An approach to teaching scientific English discourse. RELC Journal, 5.1, 27-40.

Wiseman, S. (1949). The marking of English composition in grammar school selection. British Journal of Educational Psychology, 19, 200-209.

Table 1

TOEFL Score Data for Total Sample of International Candidates and Three Language Groups

TOEFL Scores	Mean	SD	TOEFL Means*
Total Sample (N=542)			
Section I. Listening Comprehension	51.70	7.38	---
Section II. Structure and Written Express.	52.03	7.23	---
Section III. Reading Comprehension	52.26	6.68	---
Total	519.97	64.08	---
Arabic Language Group (N=138)			
Section I. Listening Comprehension	48.28	8.02	49
Section II. Structure and Written Express.	48.11	7.92	45
Section III. Reading Comprehension	47.37	6.85	45
Total	479.22	67.62	463
Chinese Language Group (N=230)			
Section I. Listening Comprehension	51.99	5.62	50
Section II. Structure and Written Express.	52.49	5.75	50
Section III. Reading Comprehension	52.80	5.33	51
Total	524.26	48.62	503
Spanish Language Group (N=174)			
Section I. Listening Comprehension	54.03	7.90	52
Section II. Structure and Written Express.	54.53	7.16	48
Section III. Reading Comprehension	55.41	5.94	51
Total	546.61	63.46	504

*TOEFL score means for the separate language groups, as reported in the _TOEFL Test and Score Manual_ (1983).

Table 2

Scores on Writing Samples, TOEFL, GRE General Test, and LSAT Writing Test for Sample of GRE Candidates

Scores	Mean	SD	N
Writing Sample Scores (International and U. S. Candidates)			
Holistic (over four topics)	14.78	4.97	172
Discourse-level (over two topics)	7.75	2.44	
Sentence-level (over two topics)	7.36	2.78	
TOEFL (International Candidates Only)			
Section I. Listening Comprehension	52.81	6.37	124
Section II. Structure & Written Expression	54.27	5.44	
Section III. Reading Comprehension	54.46	5.05	
Total	538.45	47.51	
GRE General Test (International and U. S. Candidates)			
GRE-Verbal	384.59	144.64	172
GRE-Quantitative	634.83	114.54	
GRE-Analytical	487.56	120.43	
LSAT Writing Test (U. S. Candidates Only)			
Usage Section (35 items)	21.05	6.62	43
Sentence Correction Section (25 items)	14.72	4.32	
Total (60 items)	35.77	10.25	

Table 3

Scores on Writing Samples, TOEFL, GRE General Test, and LSAT Writing Test for United States and International Samples of GRE Candidates

	International			United States		
Scores	Mean	SD	N	Mean	SD	N
Writing Sample Scores						
Holistic (over four topics)	12.56	3.30	124	20.53	3.81	48
Discourse-level (over two topics)	6.70	1.80		10.46	1.67	
Sentence-level (over two topics)	6.05	1.85		10.73	1.72	
TOEFL (International Candidates Only)						
Section I. Listening Comprehension	52.81	6.37	124			
Section II. Structure & Written Expression	54.27	5.44				
Section III. Reading Comprehension	54.46	5.05				
Total	538.45	47.51				
GRE General Test (International and U. S. Candidates)						
GRE-Verbal	320.00	91.61	124	551.46	121.29	48
GRE-Quantitative	660.81	101.07		567.71	120.91	
GRE-Analytical	450.48	98.69		583.33	119.53	
LSAT Writing Test (U. S. Candidates Only)						
Usage Section (35 items)				21.05	6.62	43
Sentence Correction Section (25 items)				14.72	4.32	
Total (60 items)				35.77	10.25	

Table 4

Scores on Writing Samples for Total Sample and International Language Groups

	Intl. Total		Arabic		Chinese		Spanish	
Writing Sample Scores	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Holistic scores								
Space	3.07	1.17	2.80	1.31	2.99	1.06	3.39	1.13
Leisure	3.22	1.15	2.85	1.22	3.20	1.09	3.55	1.08
Farming	3.19	1.17	2.91	1.12	3.02	1.14	3.64	1.12
Continents	3.14	1.09	2.91	1.15	2.98	.99	3.52	1.06
Total for four	12.63	3.89	11.47	4.18	12.20	3.58	14.11	3.62
D/S Scores								
D score—Space	3.29	1.17	2.90	1.17	3.27	1.09	3.62	1.17
S score—Space	2.89	1.20	2.57	1.28	2.81	1.07	3.26	1.22
D score--Farming	3.25	1.16	2.68	1.13	3.29	1.09	3.65	1.12
S score--Farming	2.96	1.17	2.43	1.05	2.95	1.10	3.39	1.16
	N= 542		N= 138		N= 230		N=174	

Table 5

Criteria Used to Evaluate Written Assignments
Saturday and Sunday Reader Questionnaire Responses

(in percentages of total of 50 respondents on Saturday,
51 respondents on Sunday)

	Degree of Importance							
	Low	Moderate			High	Blank	Mean	SD
Features of Written Assignments	1	2	3	4	5			
1. Correctness of punctuation/ spelling								
Prior to reading	4	22	40	24	6	4	3.1	.95
During holistic reading	12	38	38	6	2	4	2.5	.87
During D/S reading	8	18	29	29	12	4	3.2	1.14
2. Mastery of the conventions of grammar								
Prior to reading	0	0	32	42	20	6	3.9	.74
During holistic reading	0	14	50	28	4	4	3.2	.75
During D/S reading	0	6	20	39	33	2	4.0	.89
3. Quality of sentence structure								
Prior to reading	0	0	18	50	28	4	4.1	.69
During holistic reading	0	2	38	46	10	4	3.7	.69
During D/S reading	0	0	12	35	51	2	4.4	.70
4. Size of vocabulary								
Prior to reading	2	22	48	20	4	4	3.0	.84
During holistic reading	4	26	42	20	4	4	3.0	.91
During D/S reading	2	18	37	35	6	2	3.3	.90
5. Appropriateness of vocabulary usage								
Prior to reading	0	2	38	44	10	6	3.7	.70
During holistic reading	0	8	38	42	8	4	3.5	.77
During D/S reading	0	4	20	61	14	2	3.9	.70
6. Quality of paragraph organization								
Prior to reading	0	4	14	50	28	4	4.1	.78
During holistic reading	0	16	26	40	14	4	3.5	.94
During D/S reading	0	0	18	49	29	4	4.1	.70
7. Quality of overall paper organization								
Prior to reading	0	4	6	48	38	4	4.2	.76
During holistic reading	0	8	24	32	32	4	3.9	.96
During D/S reading	0	0	6	31	61	2	4.6	.61

Table 5 (continued)

	Degree of Importance							
	Low	Moderate			High	Blank	Mean	SD
Features of Written Assignments	1	2	3	4	5			
8. Quality of content								
Prior to reading	2	6	14	44	30	4	4.0	.96
During holistic reading	2	14	32	34	14	4	3.5	.99
During D/S reading	0	12	26	41	20	2	3.7	.93
9. Development of ideas								
Prior to reading	0	0	16	32	48	4	4.3	.75
During holistic reading	0	4	26	38	28	4	3.9	.86
During D/S reading	0	4	20	37	37	2	4.1	.86
10. Overall writing ability								
Prior to reading	2	0	4	34	56	4	4.5	.77
During holistic reading	0	0	8	24	66	2	4.6	.64
During D/S reading	0	0	8	29	59	4	4.5	.65
Meeting constraints of particular assignments:								
11. Student addresses topic adequately and directly								
Prior to reading	0	2	8	54	32	4	4.2	.68
During holistic reading	4	16	36	28	12	4	3.3	1.00
During D/S reading	2	6	29	35	26	2	3.8	.98
12. Student adopts a tone, attitude, or style appropriate to the audience								
Prior to reading	4	10	32	38	12	4	3.5	.99
During holistic reading	12	30	40	10	4	4	2.6	.98
During D/S reading	8	22	33	31	4	2	3.0	1.02
13. Student appropriately meets assignment requirements								
Prior to reading	2	2	18	44	30	4	4.0	.89
During holistic reading	6	18	32	32	8	4	3.2	1.00
During D/S reading	1	14	33	43	6	2	3.4	.88

Table 6

Criteria Used to Evaluate Written Assignments
Saturday Reader Questionnaire Responses
Prior to Reading Sessions

(in percentages of total of 50 respondents and
24 ESL Readers, 26 English Readers)

	Degree of Importance							
	Low	Moderate			High	Blank	Mean	SD
Features of Written Assignments	1	2	3	4	5			
1. Correctness of punctuation/ spelling								
All respondents	4	22	40	24	6	4	3.1	.95
ESL readers	4	25	29	29	4	8	3.0	1.00
English readers	4	19	50	19	8	0	3.1	.94
2. Mastery of the conventions of grammar								
All respondents	0	0	32	42	20	6	3.9	.74
ESL readers	0	0	21	42	29	8	4.1	.75
English readers	0	0	42	42	12	4	3.7	.69
3. Quality of sentence structure								
All respondents	0	0	18	50	28	4	4.1	.69
ESL readers	0	0	21	42	29	8	4.1	.75
English readers	0	0	15	58	27	0	4.1	.65
4. Size of vocabulary								
All respondents	2	22	48	20	4	4	3.0	.84
ESL readers	4	17	50	21	8	0	3.0	.78
English readers	0	27	46	19	8	0	3.1	.89
5. Appropriateness of vocabulary usage								
All respondents	0	2	38	44	10	6	3.7	.70
ESL readers	0	0	38	54	0	8	3.6	.50
English readers	0	4	38	35	19	4	3.7	.84
6. Quality of paragraph organization								
All respondents	0	4	14	50	28	4	4.1	.78
ESL readers	0	4	21	42	25	8	4.0	.84
English readers	0	4	8	58	31	0	4.2	.73
7. Quality of overall paper organization								
All respondents	0	4	6	48	38	4	4.2	.76
ESL readers	0	8	12	42	29	8	4.0	.93
English readers	0	0	0	54	46	0	4.5	.51

Table 6 (continued)

	Degree of Importance							
	Low	Moderate			High	Blank	Mean	SD
	1	2	3	4	5			
Features of Written Assignments								
8. Quality of content								
All respondents	2	6	14	44	30	4	4.0	.96
ESL readers	4	8	17	46	17	8	3.7	1.04
English readers	0	4	12	42	42	0	4.2	.82
9. Development of ideas								
All respondents	0	0	16	32	48	4	4.3	.75
ESL readers	0	0	21	50	21	8	4.0	.69
English readers	0	0	12	15	73	0	4.6	.70
10. Overall writing ability								
All respondents	2	0	4	34	56	4	4.5	.77
ESL readers	0	0	8	46	38	8	4.3	.65
English readers	4	0	0	23	73	0	4.6	.85
Meeting constraints of particular assignments:								
11. Student addresses topic adequately and directly								
All respondents	0	2	8	54	32	4	4.2	.68
ESL readers	0	4	8	58	21	8	4.0	.72
English readers	0	0	9	50	42	0	4.3	.63
12. Student adopts a tone, attitude, or style appropriate to the audience								
All respondents	4	10	32	38	12	4	3.5	.99
ESL readers	4	17	33	29	8	8	3.2	1.02
English readers	4	4	31	46	15	0	3.7	.94
13. Student appropriately meets assignment requirements								
All respondents	2	2	18	44	30	4	4.0	.89
ESL readers	0	4	21	38	29	8	4.0	.87
English readers	4	0	15	50	31	0	4.0	.92

Table 7

Criteria Used to Evaluate Written Assignments
Saturday Reader Questionnaire Responses
During Holistic Scoring

(in percentages of total of 50 respondents and
24 ESL Readers, 26 English Readers)

	Degree of Importance							
	Low	Moderate			High	Blank	Mean	SD
Features of Written Assignments	1	2	3	4	5			
1. Correctness of punctuation/ spelling								
All respondents	12	38	38	6	2	4	2.5	.87
ESL readers	17	42	38	4	0	0	2.2	.81
English readers	8	35	38	8	4	8	2.6	.92
2. Mastery of the conventions of grammar								
All respondents	0	14	50	28	4	4	3.2	.75
ESL readers	0	17	54	29	0	0	3.1	.68
English readers	0	12	46	27	8	8	3.3	.82
3. Quality of sentence structure								
All respondents	0	2	38	46	10	4	3.7	.69
ESL readers	0	0	50	42	8	0	3.6	.65
English readers	0	4	27	50	12	8	3.8	.74
4. Size of vocabulary								
All respondents	4	26	42	20	4	4	3.0	.91
ESL readers	4	17	50	29	0	0	3.0	.81
English readers	4	35	35	12	8	8	2.8	1.00
5. Appropriateness of vocabulary usage								
All respondents	0	8	38	42	8	4	3.5	.77
ESL readers	0	12	46	38	4	0	3.3	.76
English readers	0	4	31	46	12	8	3.7	.75
6. Quality of paragraph organization								
All respondents	0	16	26	40	14	4	3.5	.94
ESL readers	0	17	33	38	12	0	3.5	.93
English readers	0	15	19	42	15	8	3.6	.97
7. Quality of overall paper organization								
All respondents	0	8	24	32	32	4	3.9	.96
ESL readers	0	12	33	21	33	0	3.8	1.07
English readers	0	4	15	42	31	8	4.1	.83

Table 7 (continued)

	Degree of Importance							
	Low	Moderate			High	Blank	Mean	SD
	1	2	3	4	5			
Features of Written Assignments								
8. Quality of content								
All respondents	2	14	32	34	14	4	3.5	.99
ESL readers	4	8	46	29	12	0	3.4	.97
English readers	0	19	19	38	15	8	3.5	1.00
9. Development of ideas								
All respondents	0	4	26	38	28	4	3.9	.86
ESL readers	0	4	33	46	17	0	3.8	.79
English readers	0	4	19	31	38	8	4.1	.90
10. Overall writing ability								
All respondents	0	0	8	24	66	2	4.6	.64
ESL readers	0	0	17	25	54	4	4.4	.78
English readers	0	0	0	23	77	0	4.8	.43
Meeting constraints of particular assignments:								
11. Student addresses topic adequately and directly								
All respondents	4	16	36	28	12	4	3.3	1.00
ESL readers	4	21	42	21	12	0	3.2	1.00
English readers	4	12	31	35	12	8	3.4	1.00
12. Student adopts a tone, attitude, or style appropriate to the audience								
All respondents	12	30	40	10	4	4	2.6	.98
ESL readers	17	29	46	4	4	0	2.5	.98
English readers	8	31	35	15	4	8	2.8	.99
13. Student appropriately meets assignment requirements								
All respondents	6	18	32	32	8	4	3.2	1.00
ESL readers	8	21	25	38	8	0	3.2	1.13
English readers	4	15	38	27	8	8	3.2	.98

Table 8

Criteria Used to Evaluate Written Assignments
Sunday Reader Questionnaire Responses

During Discourse/Sentence Scoring

(in percentages of total of 50 respondents and
24 ESL Readers, 27 English Readers)

Features of Written Assignments	Degree of Importance							
	Low 1	Moderate 2	3	4	High 5	Blank	Mean	SD
1. Correctness of punctuation/ spelling								
All respondents	8	18	29	29	12	4	3.2	1.14
ESL readers	12	25	25	21	12	4	3.0	1.26
English readers	4	11	33	37	11	4	3.4	.99
2. Mastery of the conventions of grammar								
All respondents	0	6	20	39	33	2	4.0	.89
ESL readers	0	8	8	46	33	4	4.1	.90
English readers	0	4	30	33	33	0	4.0	.90
3. Quality of sentence structure								
All respondents	0	0	12	35	51	2	4.4	.70
ESL readers	0	0	8	33	54	4	3.3	.92
English readers	0	0	15	37	48	0	4.3	.73
4. Size of vocabulary								
All respondents	2	18	37	35	6	2	3.3	.90
ESL readers	4	12	38	38	4	4	3.3	.92
English readers	0	22	37	33	7	0	3.3	.90
5. Appropriateness of vocabulary usage								
All respondents	0	4	20	61	14	2	3.9	.70
ESL readers	0	8	25	50	12	4	3.7	.82
English readers	0	0	15	70	15	0	4.0	.56
6. Quality of paragraph organization								
All respondents	0	0	18	49	29	4	4.1	.70
ESL readers	0	0	21	38	33	8	4.1	.77
English readers	0	0	15	59	26	0	4.1	.64
7. Quality of overall paper organization								
All respondents	0	0	6	31	61	2	4.6	.61
ESL readers	0	0	8	25	62	4	4.6	.66
English readers	0	0	4	37	59	0	4.6	.58

Table 8 (continued)

	Degree of Importance							
	Low	Moderate			High	Blank	Mean	SD
Features of Written Assignments	1	2	3	4	5			
8. Quality of content								
All respondents	0	12	26	41	20	2	3.7	.93
ESL readers	0	17	33	38	8	4	3.4	.89
English readers	0	7	18	44	30	0	4.0	.90
9. Development of ideas								
All respondents	0	4	20	37	37	2	4.1	.86
ESL readers	0	4	33	42	17	4	3.7	.81
English readers	0	4	7	33	56	0	4.4	.80
10. Overall writing ability								
All respondents	0	0	8	29	59	4	4.5	.65
ESL readers	0	0	12	33	46	8	4.4	.73
English readers	0	0	4	26	70	0	4.7	.56
Meeting constraints of particular assignments:								
11. Student addresses topic adequately and directly								
All respondents	2	6	29	35	26	2	3.8	.98
ESL readers	4	8	29	33	21	4	3.6	1.08
English readers	0	4	30	37	30	0	3.9	.87
12. Student adopts a tone, attitude, or style appropriate to the audience								
All respondents	8	22	33	31	4	2	3.0	1.02
ESL readers	12	25	33	21	4	4	2.8	1.08
English readers	4	18	33	41	4	0	3.2	.93
13. Student appropriately meets assignment requirements								
All respondents	1	14	33	43	6	2	3.4	.88
ESL readers	4	21	29	38	4	4	3.2	.98
English readers	0	7	37	48	7	0	3.6	.75

Table 9

Reader Responses to Questions About Scoring Systems on Saturday and Sunday Questionnaires

(in whole percentages of total of 50 respondents on Saturday, 51 respondents on Sunday)

	Responses					
Questions	Yes	No	Maybe	Blank	Mean	SD
1. Is this kind of scoring appropriate to and useful in the classroom?						
Saturday	70	16	12	2	1.4	.70
Sunday	57	33	4	6	1.4	.58
3. Do you feel that the scores you were asked to give were appropriate for the papers you read in this session?						
Saturday	82	8	8	2	1.2	.60
Sunday	80	12	2	6	1.2	.43
4. After this reading experience, do you feel that it is possible to make clear distinctions between papers at adjacent score intervals?						
Saturday	60	18	8	14	1.4	.66
Sunday	45	35	18	2	1.7	.76
5. Do you feel that it would be possible to assign descriptions to each of the score intervals used...?						
Saturday	50	26	10	14	1.5	.70
Sunday	51	37	10	2	1.6	.67
Questions only on Sunday questionnaire:						
6. Are the kinds of scores we asked you to assign appropriate to the papers that were read?						
The holistic judgments?	82	10	2	6	1.1	.41
The two-score judgments?	74	16	4	6	1.2	.53
7. Regarding the two-score judgments, did you feel that they were						
Independent?	51	33	10	6	1.6	.68
Pertinent?	82	4	4	10	1.1	.45
All-inclusive?	63	28	2	8	1.3	.52
Should have been divided differently?	4	78	4	14	2.0	.30

Table 10

ESL and English Reader Responses to Questions About Scoring Systems on Saturday Questionnaire

(in whole percentages of total of 50 respondents; 26 ESL, 24 English readers)

	Responses					
Questions	Yes	No	Maybe	Blank	Mean	SD
1. Is this kind of scoring appropriate to and useful in the classroom?						
All respondents	70	16	12	2	1.4	.70
ESL readers	58	21	17	4	1.6	.79
English readers	81	12	8	0	1.3	.60
3. Do you feel that the scores you were asked to give were appropriate for the papers you read in this session?						
All respondents	82	8	8	2	1.2	.60
ESL readers	71	8	17	4	1.4	.79
English readers	92	8	0	0	1.1	.27
4. After this reading experience, do you feel that it is possible to make clear distinctions between papers at adjacent score intervals?						
All respondents	60	18	8	14	1.4	.66
ESL readers	54	21	12	12	1.5	.75
English readers	65	15	4	15	1.3	.55
5. Do you feel that it would be possible to assign descriptions to each of the score intervals used...?						
All respondents	50	26	10	14	1.5	.70
ESL readers	67	12	8	12	1.3	.66
English readers	35	39	12	15	1.7	.70

Table 11

ESL and English Reader Responses to Questions About Scoring Systems on Sunday Questionnaires

(in whole percentages of total of 51 respondents;
24 ESL, 27 English readers)

	Responses					
Questions	Yes	No	Maybe	Blank	Mean	SD
1. Is this kind of scoring appropriate to and useful in the classroom?						
All respondents	57	33	4	6	1.4	.58
ESL readers	54	33	4	8	1.5	.60
English readers	59	33	4	4	1.4	.58
3. Do you feel that the scores you were asked to give were appropriate for the papers you read in this session?						
All respondents	80	12	2	6	1.2	.43
ESL readers	79	12	0	8	1.2	.35
English readers	82	11	4	4	1.2	.49
4. After this reading experience, do you feel that it is possible to make clear distinctions between papers at adjacent score intervals?						
All respondents	45	35	18	2	1.7	.76
ESL readers	46	29	21	4	1.7	.81
English readers	44	41	15	0	1.7	.72
5. Do you feel that it would be possible to assign descriptions to each of the score intervals used...?						
All respondents	51	37	10	2	1.6	.67
ESL readers	50	33	12	4	1.6	.72
English readers	52	41	7	0	1.6	.64
Questions only on Sunday questionnaire:						
6. Are the kinds of scores we asked you to assign appropriate to the papers that were read?						
The holistic judgments?						
All respondents	82	10	2	6	1.1	.41
ESL readers	71	17	4	8	1.3	.55
English readers	93	4	0	4	1.0	.20

Table 11 (continued)

	Responses					
Questions	Yes	No	Maybe	Blank	Mean	SD
6. Are the kinds of scores we asked you to assign appropriate to the papers that were read?						
The two-score judgments?						
All respondents	74	16	4	6	1.2	.53
ESL readers	71	17	4	8	1.27	.55
English readers	78	15	4	4	1.23	.51
7. Regarding the two-score judgments, did you feel that they were						
Independent?						
All respondents	51	33	10	6	1.6	.68
ESL readers	38	42	17	4	1.8	.74
English readers	63	26	4	7	1.4	.57
Pertinent?						
All respondents	82	4	4	10	1.1	.45
ESL readers	83	0	8	8	1.2	.59
English readers	82	8	0	11	1.1	.28
All-inclusive?						
All respondents	63	28	2	8	1.3	.52
ESL readers	62	25	4	8	1.4	.58
English readers	63	30	0	7	1.3	.48
Should have been divided differently?						
All respondents	4	78	4	14	2.0	.30
ESL readers	4	75	8	12	2.0	.38
English readers	4	82	0	15	2.0	.21

Table 12

Factor Loadings Obtained from the Principal Axes Factor Analysis Seven Writing Sample and TOEFL Variables

(N=560)

Variables	Factor I Loading	Factor II Loading
Writing Samples		
Holistic score--Space	.80	.32
Holistic score--Leisure	.78	.33
Holistic score--Farming	.80	.32
Holistic score--Continents	.75	.34
TOEFL		
Section I. Listening Comprehension	.26	.87
Section II. Structure and Written Expression	.43	.79
Section III. Reading Comprehension	.41	.82

Table 13

Factor Loadings Obtained from the Principal Axes Factor Analysis Seven Writing Sample and TOEFL Variables

Arabic language group (N=139)

Variables	Factor I Loading	Factor II Loading
Writing Samples		
Holistic score--Space	.79	.37
Holistic score--Leisure	.78	.42
Holistic score--Farming	.84	.25
Holistic score--Continents	.82	.23
TOEFL		
Section I. Listening Comprehension	.19	.93
Section II. Structure and Written Expression	.58	.66
Section III. Reading Comprehension	.60	.69

(Accounting for 79% of total variance)

Table 14

Factor Loadings Obtained from the Principal Axes Factor Analysis Seven Writing Sample and TOEFL Variables

Chinese language group (N=230)

Variables	Factor I Loading	Factor II Loading
Writing Samples		
Holistic score--Space	.82	.24
Holistic score--Leisure	.81	.24
Holistic score--Farming	.82	.26
Holistic score--Continents	.72	.33
TOEFL		
Section I. Listening Comprehension	.20	.84
Section II. Structure and Written Expression	.32	.80
Section III. Reading Comprehension	.33	.84

(Accounting for 73% of total variance)

Table 15

Factor Loadings Obtained from the Principal Axes Factor Analysis Seven Writing Sample and TOEFL Variables

Spanish language group (N=191)

Variables	Factor I Loading	Factor II Loading
Writing Samples		
Holistic score--Space	.37	.74
Holistic score--Leisure	.22	.83
Holistic score--Farming	.50	.64
Holistic score--Continents	.43	.68
TOEFL		
Section I. Listening Comprehension	.82	.32
Section II. Structure and Written Expression	.80	.43
Section III. Reading Comprehension	.86	.32

(Accounting for 74% of total variance)

Table 16

Correlations of Holistic Scores, D/S Scores, and TOEFL Scores (total sample of 542 candidates)

	Holistic Scores					D/S Scores				TOEFL Scores		
	C/C		C/G			Space		Farming		I	II	III
	S	L	F	C	T	D	S	D	S	LC	SWE	RC
Holistic Compare/Contrast												
Space												
Leisure	.65											
Holistic Chart/Graph												
Farming	.65	.66										
Continents	.62	.60	.61									
Total holistic	.86	.85	.86	.82								
Discourse/Sentence												
Space—D	.74	.62	.64	.58	.76							
Space—S	.72	.61	.61	.56	.74	.83						
Farming—D	.58	.59	.72	.56	.72	59	.52					
Farming—S	.66	.65	.72	.61	.78	.63	.63	.84				
TOEFL												
I. Listening C.	.50	.53	.50	.49	.59	.52	.51	.53	.56			
II. S & W Express.	.59	.57	.60	.58	.69	.60	.60	.58	.61	.68		
III. Reading C.	.60	.58	.58	.58	.69	.60	.58	.62	.63	.72	.79	
Total	.62	.62	.62	.61	.72	.63	.62	.63	.66	.89	.91	.92

Table 17
Means and Standard Deviations for Writing Sample and TOEFL Scores

Language Group	N		Writing Sample Scores									TOEFL Scores			
			Holistic Scores					Discourse and Sentence Level Scores							
			Compare/Contrast		Chart/Graph		Total	Discourse Level		Sentence Level		Listening Comprehension	Structure & Written Expression	Reading Comprehension	Total
			Space	Recreation	Continents	Farming		Space	Farming	Space	Farming				
Arabic															
		M	2.80	2.85	2.91	2.91	11.47	2.90	2.68	2.57	2.43	48.28	48.11	47.37	479.22
	174	SD	1.31	1.22	1.15	1.12	4.14	1.17	1.13	1.28	1.05	8.02	7.92	6.85	67.62
Chinese															
		M	2.99	3.20	2.98	3.02	12.20	3.27	3.29	2.81	2.95	51.99	52.49	52.80	524.26
	230	SD	1.06	1.09	.99	1.14	3.58	1.09	1.09	1.07	1.10	5.62	5.75	5.33	48.62
Spanish															
		M	3.39	3.55	3.52	3.64	14.11	3.62	3.65	3.26	3.39	54.03	54.53	55.41	546.61
	138	SD	1.13	1.08	1.06	1.12	3.62	1.17	1.12	1.22	1.16	7.90	7.16	5.94	63.46
TOTAL															
		M	3.07	3.22	3.14	3.19	12.63	3.29	3.25	2.82	2.96	51.70	52.03	52.26	519.97
	542	SD	1.17	1.15	1.09	1.17	3.89	1.17	1.16	1.20	1.17	7.38	7.23	6.68	64.08

Table 18

Correlations of Demographic Variables with Holistic Scores, D/S Scores, and TOEFL Scores*

(total sample of 542 international candidates)

	r
Age	
Holistic score—Farming	-.15
Holistic score—Total	-.15
Discourse score—Farming	-.18
TOEFL—Section I (LC)	-.25
TOEFL—Section II (S & WE)	-.12
TOEFL—Section III (RC)	-.08 (.05)
TOEFL—Total	-.17
Sex	
TOEFL—Section I (LC)	.15

	r
Number Years of English	
Holistic score—Space	.14
Holistic score—Leisure	.15
Holistic score—Farming	.11 (.05)
Holistic score—Continents	.13
Holistic score—Total	.15
Sentence score—Farming	.16
Discourse score—Farming	.13
Sentence score—Space	.13
TOEFL—Section I (LC)	.20
TOEFL—Section II (S & WE)	.12
TOEFL—Section III (RC)	.11 (.05)
TOEFL—Total	.16

*Significant at the .01 level, unless otherwise specified

Table 19

Correlations of Demographic Variables with Holistic Scores, D/S Scores, and TOEFL Scores*

(sample of 138 Arabic language candidates)

	r
Sex	
Holistic score--Leisure	.27
Holistic score--Farming	.21 (.05)
Holistic score--Total	.24
Sentence score--Farming	.18 (.05)
TOEFL--Section I (LC)	.20 (.05)
TOEFL--Section II (S & WE)	.21 (.05)
TOEFL--Total	.20 (.05)
Major Field--Science	
Holistic score--Continents	.29
TOEFL--Section III (RC)	.30
Undergraduate Level	
Holistic score--Leisure	-.19 (.05)
Holistic score--Continents	-.22 (.05)
Holistic score--Total	-.21 (.05)
TOEFL--Section II (S & WE)	-.19 (.05)
TOEFL--Section III (RC)	-.25

	r
Number Years of English	
Holistic score--Space	.29
Holistic score--Leisure	.25
Holistic score--Farming	.27
Holistic score--Continents	.30
Holistic score--Total	.32
Discourse score--Farming	.34
Sentence score--Farming	.38
Discourse score--Space	.36
Sentence score--Space	.23
TOEFL--Section I (LC)	.26
TOEFL--Section II (S & WE)	.31
TOEFL--Section III (RC)	.35
TOEFL--Total	.35
Age	
TOEFL--Section I (LC)	-.32

*Significant at the .01 level unless otherwise specified

Table 20

Correlations of Demographic Variables with Holistic Scores, D/S Scores, and TOEFL Scores*

(sample of 230 Chinese language candidates)

	r		r
Age		**Number Years of English**	
Holistic score--Space	-.22	Holistic score--Space	.22
Holistic score--Leisure	-.25	Holistic score--Leisure	.28
Holistic score--Farming	-.24	Holistic score--Farming	.19
Holistic score--Continents	-.24	Holistic score--Continents	.26
Holistic score--Total	-.29	Holistic score--Total	.28
Discourse score--Farming	-.25	Discourse score--Farming	.17
Sentence score--Farming	-.16(.05)	Sentence score--Farming	.25
Sentence score--Space	-.17	Discourse score--Space	.15 (.05)
TOEFL--Section I (LC)	-.20	TOEFL--Section I (LC)	.18
TOEFL--Section II (S & WE)	-.25	TOEFL--Section II (S & WE)	.11 (.05)
TOEFL--Section III (RC)	-.19	TOEFL--Section III (RC)	.22
TOEFL--Total	-.24	TOEFL--Total	.20
Undergraduate Level		**Major Field--Science**	
Holistic score--Space	.23	Holistic score--Space	-.22
Holistic score--Leisure	.25	Holistic score--Leisure	-.17
Holistic score--Farming	.22	Holistic score--Total	-.18
Holistic score--Total	.26		
Discourse score--Farming	.23		
Sentence score--Farming	.26		
Discourse score--Space	.18 (.05)		
Sentence score--Space	.14 (.05)		

*Significant at the .01 level unless otherwise specified

Table 21

Correlations of Demographic Variables with Holistic Scores, D/S Scores, and TOEFL Scores*

(sample of 174 Spanish language candidates)

	r		r
Major Field--Science		Number Years of English	
Sentence score--Farming	.19 (.05)	Holistic score--Space	.30
Discourse score--Farming	.18 (.05)	Holistic score--Leisure	.33
TOEFL--Section II (S & WE)	.28	Holistic score--Farming	.39
TOEFL--Section III (RC)	.22	Holistic score--Continents	.33
		Holistic score--Total	.41
		Sentence score--Farming	.38
		Discourse score--Farming	.35
Major Field--Business		Sentence score--Space	.30
		Discourse score--Space	.35
Sentence score--Farming	-.18 (.05)	TOEFL--Section I (LC)	.55
TOEFL--Section III (RC)	-.17 (.05)	TOEFL--Section II (S & WE)	.39
		TOEFL--Section III (RC)	.41
		TOEFL--Total	.50
Age			
TOEFL--Section I (LC)	-.24	Undergraduate Level	
		TOEFL--Section II (S & WE)	-.27

*Significant at the .01 level unless otherwise specified

Table 22

Correlations of Holistic Scores, D/S Scores, TOEFL Scores, LSAT Writing Scores, and GRE Scores (sample of GRE candidates)

	Total Holistic Score	D/S Scores		TOEFL Scores				LSAT Scores			GRE Scores	
		Disc	Sent	I	II	III	Total	U	SC	Total	V	Q
Total Discourse (N=172)	.89 (.74)*											
Total Sentence (N=172)	.90 (.78)	.91 (.81)										
TOEFL (N=124)												
I. LC	(.53)	(.48)	(.52)									
II. S & WE	(.58)	(.50)	(.57)	(.50)								
III. RC	(.52)	(.54)	(.53)	(.62)	(.60)							
Total	(.64)	(.60)	(.64)	(.86)	(.82)	(.86)						
LSAT Writing (N=43)												
Usage	.38	.45	.42	—	—	—	—					
Sent. Correct.	.51	.44	.49	—	—	—	—	.75				
Total	.46	.48	.48	—	—	—	—	.96	.90			
GRE (N=172)												
Verbal	.81 (.60)	.79 (.56)	.80 (.56)	.51	.61	.72	.72	.75	.66	.76		
Quantitative	-.22 (-.15)	-.20 (-.01)	-.24 (-.06)	.04	.05	.12	.08	.18	.30	.24	-.17	
Analytical	.55 (.22)	.55 (.31)	.52 (.18)	.31	.38	.40	.42	.38	.36	.40	.62	.33

---

*Scores in parentheses are for sample of foreign candidates only

Table 23

Correlations of Holistic Scores Total* and GRE Item Type Scores

(sample of 132 cases)

				GRE Scores					
			Verbal			Quantitative		Analytical	
Scores	Hol.	SC	DV	RC	QC	M	DI	AR	LR
GRE Verbal									
Sentence Completion (SC)	.68								
Discrete Verbal (DV)	.67	.64							
Reading Comprehension (RC)	.70	.70	.64						
GRE Quantitative									
Quantitative Comparisons (QC)	-.22	-.26	-.30	-.12					
Discrete Math (M)	-.31	-.28	-.36	-.26	.76				
Data Interpretation (DI)	-.09	-.03	-.08	.00	.64	.59			
GRE Analytical									
Analytical Reasoning (AR)	.23	.15	.17	.24	.46	.35	.50		
Logical Reasoning (LR)	.64	.65	.50	.67	-.09	-.18	.02	.24	

*Holistic scores averaged over four writing samples

Table 24

GRE General Test Item Types Stepwise Regression Analysis for Holistic Score Total (N=132)

GRE Item Type Predictors	r	Standardized Regression Weight	F	R
Reading Comprehension	.70	.24	.63	.80
Discrete Verbal	.67	.25	.54	
Logical Reasoning	.64	.20	.52	
Sentence Completion	.68	.19	.62	
Data Interpretation	-.09	-.09	.50	
Analytical Reasoning	.23	.12	.39	
Mathematics	-.31	-.05	.63	
Quantitative Comparisons	-.22	-.01	.68	

Table 25

Significant Correlations of TOEFL Section I (Listening Comprehension) with Writer's Workbench Variables, TOEFL Section Scores, and Writing Sample Scores for Four Writing Sample Topics*

Writer's Workbench Text Features	Farming C/G (N=87)	Space C/C (N=81)	Continents C/G (N=82)	Leisure C/C (N=81)
Quality of development	.22 *			.22 *
Number of spelling errors	-.23 *		-.21 *	
Number to check	.24 *	.30 **	.31 **	.32 **
Variation			-.25 *	
Number of short sentences	.21 *	.34 **	.34 **	
Number of long sentences	.20 *	.30 **		
Percentage of simple sentences			-.19 *	
Percentage of complex sentences		.26 **	.21 *	
Percentage of "to be" verbs	.38 **	.22 *		
Percentage of passives	.32 **	.27 **	.22 *	.26 *
Percentage of nominalizations	-.21 *			
Number of sentences		.31 **	.28 **	.34 **
Number of words	.24 *	.45 ***	.43 ***	.46 ***
Average word length		.26 **	.30 **	
Number of questions			-.20 *	
Number of content words	.20 *	.46 ***	.42 ***	.46 ***
Percentage of content words	-.27 **			
Average length of content words		.28 *	.20 *	
Percentage of prepositions	.19 *			
Percentage of conjunctions	-.24 *			
Percentage of adverbs			.21 *	.31 **
Percentage of nouns	-.30 **			-.22 *
Kincaid readability	-.19 *			
Coleman-Liau readability		.26 **	.32 **	
Flesch readability	-.21 *			
Percentage of abstract words		.31 **		
TOEFL Section II (S & WE)	.75 ***	.67 ***	.77 ***	.80 ***
TOEFL Section III (RC)	.72 ***	.73 ***	.77 ***	.79 ***
Holistic score for topic	.57 ***	.62 ***	.62 ***	.65 ***
Sentence score for topic	.62 ***	.61 ***	(no score)	(no score)
Discourse score for topic	.62 ***	.66 ***	(no score)	(no score)

Levels of significance indicated by asterisks:=.05, **=.01, ***=.001

Table 26

Significant Correlations of TOEFL Section II (Structure and Written Expression) with Writer's Workbench Variables, TOEFL Section Scores, and Writing Sample Scores for Four Writing Sample Topics*

Writer's Workbench Text Features	Farming C/G (N=87)	Space C/C (N=81)	Continents C/G (N=82)	Leisure C/C (N=81)
Quality of development	.22 *	-.21 *		.26 *
Number of spelling errors	-.38 **	.27 **	-.36 **	-.29 **
Percentage of vague words		-.30 **		
Number to check	.30 **	.28 **	.28 **	.28 **
Variation		-.25 *	-.27 **	
Average sentence length	-.26 **	-.26 **		
Number of short sentences	.27 **	.32 **		
Number of long sentences	.27 **	.38 **		
Percentage of "to be" verbs	.46 **	.20 *		
Percentage of passives	.30 **	.40 ***		.27 **
Percentage of nominalizations			.24 *	.24 *
Number of sentences	.23 *	.37 **	.24 *	.35 **
Number of words	.27 **	.39 **	.36 **	.36 **
Average word length	.19 *	.38 **	.44 ***	.28 **
Number of questions	.24 *			
Number of imperatives			-.26 **	
Number of content words	.24 *	.42 ***	.37 **	.40 ***
Percentage of content words	-.21 *	.39 **		
Average length of content words	.34 **		.34 **	.31 **
Percentage of prepositions	.26 **			
Percentage of conjunctions	-.31 **			
Percentage of adverbs			.21 *	.24 *
Percentage of nouns	-.19 *			
Percentage of pronouns				-.24 *
Kincaid readability	-.30 **	-.23 *		
Auto readability	-.27 **	-.23 *		-.19 *
Coleman-Liau readability	-.22 *	.33 **	.46 ***	.27 **
Flesch readability				.22 *
Percentage of abstract words	.23 *	.24 *	-.18 *	
TOEFL Section I (LC)	.75 ***	.67 ***	.77 ***	.80 ***
TOEFL Section III (RC)	.34 ***	.83 ***	.79 ***	.86 ***
Holistic score for topic	.67 ***	.62 ***	.69 ***	.64 ***
Sentence score for topic	.70 ***	.68 ***	(no score)	(no score)
Discourse score for topic	.65 ***	.72 ***	(no score)	(no score)

Levels of significance indicated by asterisks:=.05, **=.01, ***=.001

Table 27

Significant Correlations of TOEFL Section III (Reading Comprehension) with Writer's Workbench Variables, TOEFL Section Scores, and Writing Sample Scores for Four Writing Sample Topics*

Writer's Workbench Text Features	Farming C/G (N=87)	Space C/C (N=81)	Continents C/G (N=82)	Leisure C/C (N=81)
Quality of development		-.23 *		.21 *
Number of suggestions to substitute	-.21 *			
Number of spelling errors	-.32 **	.27 **	-.28 **	-.23 *
Number to check	.34 **	.29 **	.36 **	.28 **
Number of punctuation errors		-.18 *		-.22 *
Variation	-.19 *		-.24 *	
Average sentence length	-.18 *	-.22 *		
Number of short sentences	.22 *	.30 **	.21 *	
Number of long sentences	.23 *	.32 **		
Percentage of "to be" verbs	.39 **	.27 **		
Percentage of passives	.28 **	.43 ***		.31 **
Percentage of nominalizations		.23 **		.29 **
Number of sentences		.28 **		.33 **
Number of words	.25 **	.34 **	.35 **	.37 **
Average word length	.19 *	.44 ***	.44 ***	.41 ***
Number of questions	.18 *	.36 **	-.27 **	
Number of imperatives			-.26 **	
Number of content words	.22 *		.36 **	.41 ***
Percentage of content words	.21 *			.18 *
Average length of content words	.37 **	.49 ***	.33 **	.43 ***
Percentage of prepositions	.36 **			
Percentage of conjunctions	-.36 **		-.20 *	
Percentage of adverbs				.24 *
Percentage of nouns				-.35 **
Percentage of pronouns		-.24 *		
Kincaid readability	-.21 *			
Auto readability	-.19 *			
Coleman-Liau readability			.47 ***	.42 ***
Flesch readability				.35 **
Percentage of abstract words	.21 *	.33 **	-.22 *	
TOEFL Section I (LC)	.72 ***	.73 ***	.77 ***	.79 ***
TOEFL Section II (S & WE)	.84 ***	.83 ***	.79 ***	.86 ***
Holistic score for topic	.62 ***	.65 ***	.63 ***	.61 ***
Sentence score for topic	.66 ***	.61 ***	(no score)	(no score)
Discourse score for topic	.63 ***	.67 ***	(no score)	(no score)

---

*Levels of significance indicated by asteriks: *=.05, **=.01, ***=.001

Table 28

Significant Correlations of Holistic Scores on Writing Samples with Writer's Workbench Variables, TOEFL Section Scores, and Writing Sample Scores for Four Writing Sample Topics*

Writer's Workbench Text Features	Farming C/G (N=87)	Space C/C (N=81)	Continents C/G (N=82)	Leisure C/C (N=81)
Quality of development	.31 **	-.20 *		.26 **
Number of spelling errors	-.28 **	-.35 **	-.26 **	-.26 **
Number to check	.31 **	.37 **	.33 **	.41 ***
Number of punctuation errors		-.20 **		
Variation		-.19 *		
Average sentence length		-.21 *		
Number of short sentences	.46 **	.43 ***		.31 **
Number of long sentences	.37 **	.37 **		.25 *
Percentage of simple sentences		.23 *		
Percentage of "to be" verbs	.33 **	.30 **	-.19 *	
Percentage of passives	.32 **	.41 ***		.24 *
Percentage of nominalizations	.20 *		-.18 *	
Number of sentences	.37 **	.45 ***	.28 **	.58 ***
Number of words	.49 **	.56 ***	.47 ***	.66 ***
Average word length	.22 *	.28 **	.34 **	
Number of questions	.19 *			
Number of content words	.46 **	.60 ***		.69 ***
Percentage of content words	-.27 *		.46 ***	.19 *
Average length of content words	.26 *	.37 **	.30 **	
Percentage of prepositions	.38 **	.26 **		.19 *
Percentage of conjunctions	-.21 *	-.21 *		-.22 *
Percentage of adverbs				.29 **
Percentage of nouns	-.21 *			
Percentage of adjectives			-.19 *	
Kincaid readability	-.19 *	-.24 *		
Auto readability		-.23 *		
Coleman-Liau readability			.39 **	
Percentage of abstract words		.20 *	-.23 *	
TOEFL Section I (LC)	.57 ***	.62 ***	.62 ***	.65 ***
TOEFL Section II (S & WE)	.67 ***	.62 ***	.69 ***	.64 ***
TOEFL Section III (RC)	.62 ***	.65 ***	.63 ***	.61 ***
Sentence score for topic	.70 ***	.82 ***	(no score)	(no score)
Discourse score for topic	.79 ***	.83 ***	(no score)	(no score)

Levels of significance indicated by asterisks:=.05, **=.01, ***=.001

Table 29

Significant Correlations of Discourse/Sentence Scores on Writing Samples with Writer's Workbench Variables, TOEFL Section Scores, and Writing Sample Scores for Four Writing Sample Topics*

	Sentence-level		Discourse-level	
Writer's Workbench Text Features	Farming C/G (N=87)	Space C/C (N=81)	Continents C/G (N=82)	Leisure C/C (N=81)
Quality of development	.26 **	-.20 *	.28 **	-.23 *
Number of spelling errors	-.38 **	-.29 **	-.26 **	-.26 **
Percentage of vague words		-.19 *		-.29 **
Number to check	.25 *	.24 *	.27 **	.44 ***
Number of punctuation errors		-.22 *		
Variation		-.21 *		
Average sentence length	-.18 *	-.27 **		-.22 *
Number of short sentences	.23 *	.40 ***	.24 *	.45 ***
Number of long sentences	.31 **	.37 **	.28 **	.35 **
Percentage of simple sentences		.21 *		.22 *
Percentage of "to be" verbs	.42 **	.28 **	.40 ***	.25 *
Percentage of passives	.32 **	.36 **	.33 **	.32 **
Number of sentences	.22 *	.44 **	.20 *	.50 ***
Number of words	.27 **	.48 ***	.34 **	.58 ***
Average word length	.31 **	.32 **	.34 **	.32 **
Number of content words	.24 *	.52 ***	.30 **	.64 ***
Percentage of content words	-.27 *		-.28 **	
Average length of content words	.41 ***	.37 **	.42 ***	.37 **
Percentage of prepositions	.38 ***	.22 *	.36 **	.19 *
Percentage of conjunctions	-.29 **	-.21 *	-.22 *	
Percentage of nouns	-.23 *		-.23 *	
Kincaid readability		-.26 **		-.26 **
Auto readability		-.25 **		-.25 *
Coleman-Liau readability		-.29 **		.29 *
Percentage of abstract words		.29 **		
TOEFL Section I (LC)	.62 ***	.61 ***	.62 ***	.66 ***
TOEFL Section II (S & WE)	.70 ***	.68 ***	.65 ***	.72 ***
TOEFL Section III (RC)	.66 ***	.61 ***	.63 ***	.67 ***
Holistic score for topic	.70 ***	.82 ***	.79 ***	.83 ***
Sentence score for topic			.84 ***	.84 ***
Discourse score for topic	.84 ***	.84 ***		

Levels of significance indicated by asterisks:=.05, **=.01, ***=.001

Table 30

Writer's Workbench Stepwise Regression Analyses for TOEFL Section II. Structure and Written Expression

Space Topic (Compare/Contrast)

Independent Variables Writers's Workbench	Standardized Regression Weight	T Statistic	Mean	SD
Number of content words	.44	4.63	108.99	36.45
Average length of content words	.56	.31	5.89	.39
Number of suggestions-other	-.28	-3.28	.01	.11
Number of suggestions to omit	-.27	-2.94	1.55	1.44
Flesch readability formula	-.42	-3.39	1.12	.28
Coleman readability formula	.66	2.92	2.63	.31
Percentage of nouns	-.21	-2.10	9.15	8.70

R^2= .70, standard error= 6.05, N= 81
Mean for TOEFL Section II= 52.32, SD= 8.07

Recreation Topic (Compare/Contrast)

Independent Variables Writers's Workbench	Standardized Regression Weight	T Statistic	Mean	SD
Number of content words	.50	5.27	108.48	40.87
Number of spelling errors	-.39	-4.21	4.88	5.01
Percentage of nominalizations	.21	2.28	1.07	1.07
Number of suggestions to substitute	-.21	-2.35	.84	1.36
Number of questions	-.19	-2.05	.07	.41

R^2= .63, standard error= 6.42, N= 81
Mean for TOEFL Section II= 53.27, SD= 7.95

Farming Topic (Chart/Graph)

Independent Variables Writers's Workbench	Standardized Regression Weight	T Statistic	Mean	SD
Percentage of "to be" verbs	.30	3.35	72.10	20.12
Number of spelling errors	-.26	-3.04	4.51	4.18
Number of short sentences	.24	2.91	2.46	1.57
Length of content words	.28	2.93	5.74	.48
Flesch readability formula	-.26	-2.88	1.02	.29
Percentage of prepositions	.17	2.04	1.29	.31

R^2= .71, standard error= 6.00, N= 87
Mean for TOEFL Section II= 53.43, SD= 8.15

Continents Topic (Chart/Graph)

Independent Variables Writers's Workbench	Standardized Regression Weight	T Statistic	Mean	SD
Coleman readability formula	.40	5.05	.93	.19
Number of words	.46	5.63	189.71	67.85
Number of spelling errors	-.36	-4.38	4.30	3.61
Percentage of abstract words	-.21	-2.58	.12	.11

R^2= .72, standard error= 5.76, N= 82
Mean for TOEFL Section II= 52.83, SD= 8.02

Table 31

Writer's Workbench Stepwise Regression Analyses for Holistic Scores

Independent Variables Writers's Workbench	Standardized Regression Weight	T Statistic	Mean	SD
Space Topic (Compare/Contrast)				
Number of content words	.63	7.63	108.99	36.45
Number of spelling errors	-.25	-3.22	5.33	3.72
Average length of content words	.26	3.30	5.89	.39
Number of suggestions to omit	-.18	-2.19	1.55	1.44
R^2= .74, standard error= 1.01, N= 81 Mean of Space holistic scores= 3.23, SD= 1.47				
Recreation Topic (Compare/Contrast)				
Number of content words	.73	10.64	108.48	40.87
Number of spelling errors	-.40	-5.77	4.88	5.01
Percentage of conjunctions	-.16	-2.35	.47	.16
R^2= .81, standard error= .91, N= 81 Mean of Recreation holistic scores= 3.39, SD= 1.50				
Farming Topic (Chart/Graph)				
Number of words	.45	5.41	193.79	72.46
Number of spelling errors	-.28	-3.56	4.51	4.18
Percentage of prepositions	.28	3.51	1.29	.31
Number of long sentences	.26	3.17	1.02	.84
Average length of content words	.17	2.11	5.74	.48
Percentage of sentence beginnings	.16	2.06	62.61	19.47
R^2= .75, standard error= .97, N= 87 Mean of Farming holistic scores= 3.36, SD= 1.41				
Continents Topic (Chart/Graph)				
Number of words	.56	7.16	189.71	67.85
Coleman readability formula	.38	4.90	.93	.19
Percentage of abstract words	-.26	-3.32	.12	.11
Number of spelling errors	-.27	-3.46	4.30	3.61
Percentage of adjectives	-.17	-2.15	1.64	.44
R^2= .76, standard error= .98, N= 82 Mean of Continents holistic scores= 3.30, SD= 1.43				

Table 32

Writer's Workbench Stepwise Regression Analyses for D/S Scores

Independent Variables Writers's Workbench	Standardized Regression Weight	T Statistic	Mean	SD
Space Topic (Compare/Contrast)				
For Discourse-level scores:				
Number of content words	.66	8.23	108.99	36.45
Average length of content words	.40	4.93	5.89	.39
Number of suggestions to subsitute	-.21	-2.57	2.01	1.65
Flesch readability formula	-.20	-2.41	1.12	.28
R^2= .76, standard error= .89, N= 81 Mean of Space Discourse scores= 3.24, SD= 1.34				
For Sentence-level scores:				
Number of content words	.46	5.43	108.99	36.45
Average length of content words	.41	4.50	5.89	.39
Flesch readability formula	-.29	-3.12	1.12	.28
Punctuation	-.20	-2.33		
R^2= .68, standard error= .94, N= 87 Mean of Space Sentence scores= 2.84, SD= 1.25				
Farming Topic (Chart/Graph)				
For Discourse-level scores:				
Average length of content words	.61	5.52	5.74	.48
Number of words	.25	2.95	193.79	72.46
Coleman readability formula	-.33	-2.87		
Percentage of prepositions	.24	2.77	1.29	.31
Percentage of nouns	-.33	-3.46	2.80	.28
Percentage of pronouns	-.23	-2.31	.56	.26
R^2= .71, standard error= .85, N= .82 Mean of Farming Discourse scores= 3.37, SD= 1.16				
For Sentence-level scores:				
Percentage of "to be" verbs	.28	3.61	72.10	20.12
Percentage of prepositions	.27	3.50	1.29	.31
Number of long sentences	.34	4.62	1.02	.84
Number of spelling errors	-.23	-2.90	4.51	4.18
Average length of content words	.29	3.39	5.74	.48
Percentage of nominalizations	.22	2.80	2.80	1.33
R^2= .75, standard error= 8.30 Mean of Farming Sentence scores= 3.04, SD= 1.21				

Appendixes

A. Writing Assessment Test Instructions and Topics

B. List of Readers for Reading Weekend
   List of Subject Matter Readers

Appendix A

Writing Assessment Test Instructions and Topics

TOEFL Writing Sample

Total Time - 2 hours
4 Topics
30 Minutes Per Topic

Please PRINT the following information:

Name: ______________________ ______________________ __________
Family Name First Name M.I.

TOEFL APPLICATION NUMBER ☐☐☐☐☐☐☐

Native Country: ______________________

What major subject do you plan to study? ______________________

How many years have you studied English? ______________________

Please CHECK the appropriate boxes:

Applying for admission as:
☐ Undergraduate student
☐ Graduate student

Sex:
☐ Male
☐ Female

General Instructions

You will have two hours to plan and write essays on the four topics in this booklet. At the end of each thirty-minute period, the supervisor will tell you to stop writing on one topic and begin writing on the next topic. These topics are presented to give you an opportunity to show how well you can write. There are many possible responses to each topic but no "right" answers. What is important, therefore, is that you take care to express your thoughts on each topic clearly and effectively. How well you write is more important than how much you write. However, to cover each topic adequately, you should write more than one paragraph.

Write your essays in this booklet, using the lined pages that follow each topic. You will have enough space if you write on every line, avoid wide margins, and keep your handwriting to a reasonable size. You may use the space immediately below each topic to make notes, if you wish.

PLEASE DO NOT OPEN THIS BOOKLET UNTIL YOU ARE TOLD TO DO SO.

Unauthorized reproduction of all or portions of this booklet is prohibited.
Copyright owned by Educational Testing Service, Princeton, NJ 08541.

TIME - 30 MINUTES

Some people say that exploration of outer space has many advantages; other people feel that it is a waste of money and other resources. Write a brief essay in which you discuss each of these positions. Give one or two advantages and disadvantages of space exploration, and explain which position you support.

THIS SPACE MAY BE USED FOR NOTES.

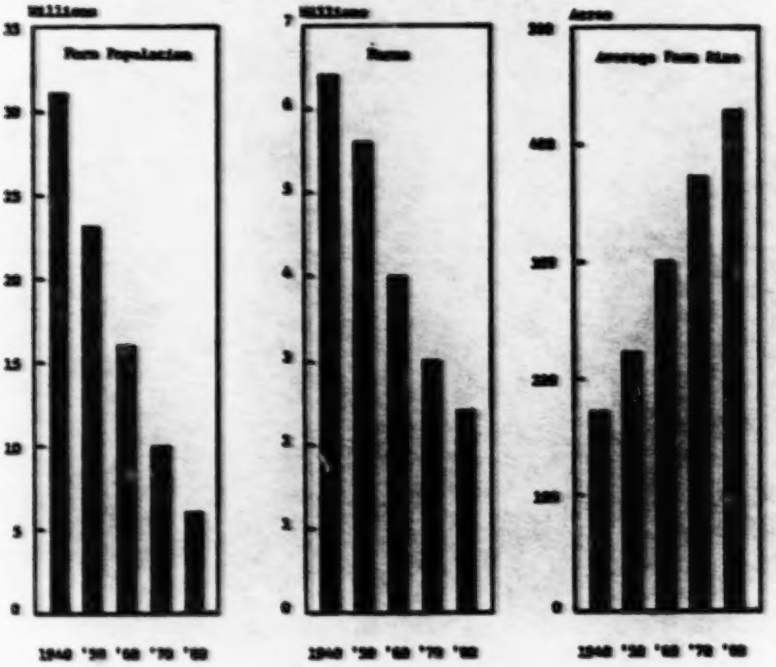
TIME - 30 MINUTES

Many people enjoy active physical recreation like sports and other forms of exercise. Other people prefer intellectual activities like reading or listening to music. In a brief essay, discuss one or two benefits of physical activities and of intellectual activities. Explain which kind of recreation you think is more valuable to someone your age.

THIS SPACE MAY BE USED FOR NOTES.

TIME - 30 MINUTES

CHANGES IN FARMING IN THE U.S.: 1940 - 1980



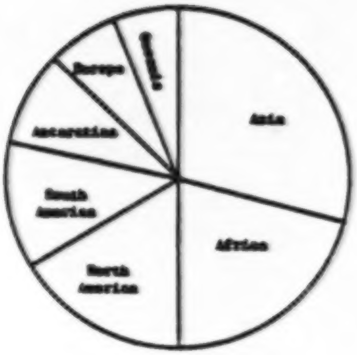
Suppose that you are writing a report in which you must interpret the three graphs shown above. Write the section of that report in which you discuss how the graphs are related to each other and explain the conclusions you have reached from the information in the graphs. Be sure the graphs support your conclusions.

THIS SPACE MAY BE USED FOR NOTES.

TIME - 30 MINUTES

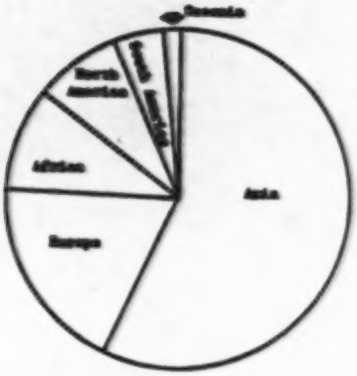
AREA AND POPULATION OF CONTINENTS

AREA



Asia	30%
Africa	20%
North America	16%
South America	12%
Antarctica	9%
Europe	7%
Oceania	6%

POPULATION



Asia	58%
Europe	16%
Africa	11%
North America	9%
South America	5%
Oceania	1%
Antarctica	0%

Suppose you are to write a report in which you interpret these charts. Discuss how the information in the Area chart is related to the information in the Population chart. Explain the conclusions you have reached from the information in the two charts. Be sure the charts support your conclusions.

YOU MAY USE THIS SPACE FOR NOTES.

Appendix B

List of Readers for Reading Weekend
List of Subject Matter Readers

Essay Reading Participants
January 28-29, 1984

Name	Affiliation	State	Subject
	CHIEF READERS		
McNamara, Susan	William Paterson College	NJ	English
Walling, William	Rutgers University	NJ	English
	TABLE LEADERS		
Arena, Louis	University of Delaware	DE	ESL
Berezovsky, Helen	University of Pennsylvania	PA	ESL
Dyer, Patricia	University of Delaware	DE	ESL
Earisman, Delbert	Upsala College	NJ	English
Lorenzi, Robert	Camden County College	NJ	English
Morgan, RoseAnn	Middlesex County College	NJ	English
Olson, Jerry	Middlesex County College	NJ	English
Taylor, Barry	University of Pennsylvania	PA	ESL
	ESL READERS		
Baron, Melvyn	Kingsborough Comm. College	NY	ESL
Carew, Pat	Nyack High School	NY	ESL
Carty, Kathleen	Columbia University	NY	ESL
David, Elizabeth	Princeton Adult Education	NJ	ESL
Della Torre, Thomas	Bergen Community College	NJ	ESL
Emery, Cornelia	Rater for Test of Spoken English	PA	ESL
Haffar, Shirley	SUNY-New Paltz	NY	ESL
Halliday, Cynthia	SUNY-New Paltz	NY	ESL
Hay, Alice	Pennington School	NJ	ESL
Lebowitz, Regina	N.Y.C. Technical College	NY	ESL
Lunt, Ruth	Rutgers University	NJ	ESL
McDowell, Alfred	Bergen Community College	NJ	ESL
Reilly, Joseph	Brooklyn Polytech	NY	ESL
Ruiz, Aida	Hostos Community College	NY	ESL
Sayre, Johanna	SUNY-New Paltz	NY	ESL
Shanefield, Elizabeth	Princeton Adult School	NJ	ESL
Slighton, Margaret	Private ESL Tutor	NJ	ESL
Stansfield, Charles	TOEFL staff	NJ	ETS
Stewart-Ghali, Denise	SUNY-New Paltz	NY	ESL
Suomi, Barbara	TOEFL staff	NJ	ETS
Tolo, Marc	Pennington School	NJ	ESL
Van Duren, David	Bergen Community College	NJ	ESL
Villaneuva, Alfredo	Hostos Community College	NY	ESL

ENGLISH READERS

Name	Affiliation	State	Subject
Asher, Deborah	Union County College	NJ	English
Billiar, Donald	Union County College	NJ	English
Buscemi, Santi	Middlesex County College	NJ	English
Cirasa, Robert	Kean College of New Jersey	NJ	English
Collins, John	Glassboro State College	NJ	English
Collins, Marilyn	Glassboro State College	NJ	English
Conlon, Michael	William Paterson College	NJ	English
Daniels, Barbara	Camden County College	NJ	English
Edge, Donald	Camden County College	NJ	English
Granger, Virgie	Wm. Paterson College	NJ	English
Gruenberg, Diane	Edison State College	NJ	English
King, Barbara	Rutgers University	NJ	English
Lees, Irene	Felician College	NJ	English
Lutz, William	Rutgers University	NJ	English
Mehlman, Robert	Trenton State College	NJ	English
O'Day, Daniel	Kean College of New Jersey	NJ	English
Oszmanski, Pat	Department of Higher Education	NJ	
Otten, Ted	Mercer County Community College	NJ	English
Palladino, Mary	Glassboro State	NJ	English
Palmere, Martha	West Windsor-Plainsboro H.S.	NJ	English
Piltch, Ziva	Pace University	NJ	English
Rehbein, Edith	Middlesex County College	NJ	English
Shea, Michael	Mercer County Comm. College	NJ	English

Subject Matter Readers

Dr. Robert Stover
Department of Political Science
University of Colorado

Dr. Melvin Oliver
Department of Sociology
UCLA

Dr. Terry Lenz
Chemical Engineering Department
Colorado State University

Dr. Robert Hunsperger
Electrical Engineering Department
University of Delaware

Dr. John Trowbridge
Civil Engineering Department
University of Delaware

Dr. Gene Chesson
Civil Engineering Department
University of Delaware

Dr. Douglas Kleiber, Ph.D
Leisure Behavior Research Laboratory
Champaign, IL

Dr. Richard Fisher
Department of Education
Colorado State University